NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

Predictive Modelling: Flight Delays and Associated Factors

Hartsfield–Jackson Atlanta International Airport

Inês Viana Feiteira

Project Work report presented as partial requirement for obtaining the Master's degree in Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

PREDICTIVE MODELLING: FLIGHT DELAYS AND ASSOCIATED FACTORS

Hartsfield–Jackson Atlanta International Airport

By

Inês Viana Feiteira

Project Work report presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Knowledge Management and Business Intelligence.

Advisor: Professor Doctor Roberto Henriques, NOVA IMS

February 2018

ACKNOWLEDGEMENTS

Success is a science; if you have the conditions, you get the result.

– Oscar Wilde (1854-1900)[1]

The success here implicit was only possible with the help of fabulous people who, during these months, supported and helped me. It would not make any sense to submit this work without first acknowledging these people and expressing how grateful I am to them.

I would like to thank my supervisor, Professor Roberto Henriques, whom I have known since graduation, and whom I have always admired for his vast knowledge, competence, and professionalism. Thank you for the teachings you gave me from the beginning, either from this thesis or my journey through this university and for always been available to answer my doubts guiding me and suggesting ways to lead the development of this work.

To my parents, Ana Lúcia and Vitor, and sister, Joana, for among all people, be the ones who, at the end of the day, have to listen to all my dilemmas and concerns. For being always available, encouraging me to go further and give my all without fear. For giving me comfort in the hardest hours. For that, for providing me with everything I ever needed, and for helping me in finding my own way in life, I must be grateful.

To my love one, Diogo Ferreira, for being my partner, my best friend, and my boyfriend. Who always helped me to see through the difficulties. Who always put the highest trust in me, never letting me down. For his effort to help me find solutions that sometimes seemed not to exist. For the endless hours he has heard me speaking of things that, for him, may make no sense. For always being by my side and believing me. For everything, and more grateful than words can demonstrate, thank you.

At last, but not least, to my dearest friends - Marta Galvão, Catarina Andrade and Tiago Costa -, that even in the worst moments made me smile and provided me the best moments making it tolerable. For always being on my side in the longest hours of this journey sharing experiences and knowledge, but also for distracting me when I needed to.

[1] Oscar Fingal O'Flahertie Wills Wilde was a famous Irish dramatist, poet and novelist (for more information go to http://www.woopidoo.com/biography/oscar-wilde/)

RESUMO

Atualmente, um ponto negativo nas viagens de avião são os atrasos que, constantemente, são anunciados aos passageiros resultando numa diminuição da sua satisfação enquanto clientes. Este e outros fatores fazem com que elevados custos, tanto quantitativos como qualitativos sejam imputados às companhias. Consequentemente, existe a necessidade de prever e mitigar a existência de atrasos aéreos que pode ajudar as companhias aéreas bem como aeroportos a melhorar a sua performance e a aplicar algumas medidas, dirigidas ao consumidor, que permitiam atenuar ou até anular o efeito que estes atrasos provoca nos seus passageiros.

Deste modo, este estudo tem como principal objetivo prever a ocorrência de atrasos nas chegadas ao aeroporto internacional de Hartsfield-Jackson. Esta estimativa será possível através da elaboração de um modelo preditivo, recorrendo a diversas técnicas de *Data Mining*. Com a aplicação destas técnicas, foi possível identificar as variáveis que mais contribuíram para a existência do atraso.

No desenvolvimento deste trabalho, foi seguida a metodologia da descoberta de conhecimento em base de dados (conhecida em inglês por *Knowledge Discovery Database*, KDD). Fases como a recolha dos dados, a aplicação de técnicas de amostragem (SMOTE e Undersampling), a partição dos dados em treino e teste, o pré-processamento (dados omissos e *outliers*) e transformação dos dados (normalização dos dados e seleção de atributos), a definição de modelos a treinar (*Decision Trees*, *Random Forest* e *Multilayer Perceptron*) bem como a avaliação da performance dos modelos através de métricas variadas foram aplicadas.

Depois de testar diferentes abordagens, concluiu-se que o melhor modelo é alcançado com as variáveis relacionadas com a partida, usando o algoritmo *Multilayer Perceptron* e aplicando a técnica de SMOTE para lidar com dados não balanceados, removendo *outliers* e selecionando dez variáveis usando *GainRatio*. Por outro lado, quando as variáveis com informação da partida são excluídas, o algoritmo que melhor se destaca é o *Multilayer Perceptron* usando a técnica SMOTE, mas desta vez, incluindo os *outliers* e com quinze variáveis selecionadas novamente pelo *GainRatio*.

Em ambas as hipóteses, as variáveis explicativas que mais contribuem para a existência do atraso na chegada são relacionadas com o clima, com as características do avião e com a propagação do atraso.

Os resultados do algoritmo de *Random Forests* mostraram melhor desempenho, em relação à precisão, em comparação com outros autores (Belcastro, Marozzo, Talia, & Trunfio, 2016; Choi, Kim, Briceno, & Mavris, 2016). Contrariamente, o algoritmo *Multilayer Perceptron*, apresentou menor precisão em comparação com outro estudo equivalente (Y. J. Kim, Choi, Briceno, & Mavris, 2016).

PALAVRAS-CHAVE

Data Mining; Análise Preditiva; Atraso Aéreo; Aeroporto Internacional de Hartsfield–Jackson; Aeroporto Internacional de Atlanta.

ABSTRACT

Nowadays, a downside to traveling is the delays that are constantly advertised to passengers resulting in a decrease in customer satisfaction. These delays associated with other factors can cause costs, both quantitative and qualitative. Consequently, there is a need to anticipate and mitigate the existence of airborne delays that can help airlines and airports improving their performance or even take some consumer-oriented measures that can undo or attenuate the effect that these delays have on their passengers.

This study has as primary objective to predict the occurrence of arrival delays of the international airport of Hartsfield-Jackson. It was possible by building a predictive model, applying several Data Mining techniques. With these applications, it was possible to show the variables, among the proposals, that most contributed to the existence of the delay.

In this work, the Knowledge Discovery Database (KDD) methodology was followed. Phases such as data collection; sampling techniques (SMOTE and Undersampling); Data partitioning in training and testing; Pre-processing (missing data and outliers) and data transformation (data normalization and attribute selection); And, finally the definition of models to be trained (Decision Trees, Random Forests, and Multilayer Perceptron), as well as the evaluation of the performance of the models through varied metrics, were used.

After testing different approaches, it was concluded that the best model is achieved with the variables related to departure, using the Multilayer Perceptron algorithm and applying SMOTE to deal with unbalanced data, removing outliers and selecting ten variables using GainRatio.

On the other hand, when the variables with information of the departure are excluded, the algorithm that performs best is also the Multilayer Perceptron using the SMOTE technique but, this time, including the outliers and with fifteen variables selected again by the GainRatio.

On both hypotheses, the explanatory variables that most contributed to the existence of the delay in arrivals were related to the weather, the airplane characteristics and the propagation of the delay.

Our results for the Random Forests algorithm shown better performance, regarding accuracy, compared to other authors (Belcastro et al., 2016; Choi et al., 2016). Contrary, for the Multilayer Perceptron algorithm, was presented a lower accuracy compared to another equivalent study (Y. J. Kim et al., 2016).

KEYWORDS

Data Mining; Predictive Analysis; Flight Delays; Hartsfield–Jackson International Airport; Atlanta International Airport.

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS AND ACRONYMS

ACPD	Aviation Consumer Protection Division
AND	Airport Network Delay Model
ASOS	Automated Surface Observing System
BN	Bayesian Network
BTS	Bureau of Transportation Statistics
DT	Decision Trees
DM	Data Mining
DOT	U.S. Department of Transportation
DPA	Delay Propagation Algorithm
EDA	Exploratory Data Analysis
EM	Expectation-Maximization
EVF	Extreme Value Factor
FAA	Federal Aviation Administration
GDP	Ground Delay Program
GDP	Gross Domestic Product
GIGO	Garbage In Garbage Out
IATA	International Air Transport Association
IEM	Iowa Environmental Mesonet
IQR	Interquartile Ranges
KDD	Knowledge Discovery Database
KNN	K-Nearest Neighbor
LSTM RNN	Long Short-Term Memory Recurrent Neural Networks
METAR	Meteorological Terminal Air Report
ML	Machine Learning
MLP	Multilayer Perceptron Neural Network
NAA	National Aviation Authority

NAS	National Airspace System
NWS	National Weather Service
OAEP	Office of Aviation Enforcement and Proceedings
OF	Outlier Factor
OPM	Office of Personnel Management
OR	Operational Research
PM	Probabilistic Models
QE	Queueing Engine theory
RF	Random Forest
RITA	Research and Innovative Technology Administration
RL	Reinforcement Learning
SA	Statistical Analysis
SAS	Statistical Analysis System
SMOTE	Synthetic Minority Over-Sampling Technique
SQL	Structured Query Language
SVM	Support Vector Machine
USA	United States of America
Weka	Waikato Environment for Knowledge Analysis
WITI	Weather Impacted Traffic Index
WITI-FA	Weather Impacted Traffic Index – Forecast Accuracy
VBA	Visual Basic for Applications

1. INTRODUCTION

Inside of every problem lies an opportunity.

– Robert Kiyosaki[2]

1.1. CONTEXT AND RELEVANCE

The airline industry has grown over the years, approximately 5% per year over the last 30 years (Belobaba, Odoni, & Barnhart, 2009). Its demand grew steadily with a global passenger air traffic growth of 6.5% in 2015, far above of the 10-year average annual growth of 5.5% (International Air Transport Association (IATA), 2015). However, a high demand does not only translate into success since it can diminish the capacity to respond to a possible rupture (Rebollo & Balakrishnan, 2014). Smallen (2016) reported, for the year of 2015, the passage of 895.5 million passengers, traveling on domestic and international flights to or from the United States of America (USA), and 9 526 flights.

As a result of the volume is the congestion of the system caused by the disproportional growth between flights and airport capacity. The U.S. Department of Transportation (DOT) (2004) define the airport capacity as the number of departures and arrivals per hour that an airport can sustain with security.

This volume, together with several factors, makes two possible scenarios. In the best scenario, is possible to exchange aircraft between flights or to request aircrafts that are not currently being used. In the worst scenario, could result the cancellation and/or delay of the flight (Jarrah, Yu, Krishnamurthy, & Rakshit, 1993).

As factors exists the mechanical problems, atmospheric conditions and air traffic control issues. Delay propagation is also a factor for delays. It is defined by AhmadBeygi, Cohn, Guan, & Belobaba (2008) as a flight delay on arrival that originate flight delay in departures when a scheduled flight depends on a specific airplane or cabin crew.

Additionally, factors such as the scarcity of labor in sectors that can cause delays (holds, mechanics, etc.) and the changes in traffic management using initiatives such as Ground Delay Programs (GDP), promote delays. GDP is seen as initiatives that represent traffic management procedures where airplanes are delayed at their departure airport to manage demand and capacity at their arrival airport (Yablonsky et al., 2014).

Therefore, it becomes necessary to realize what is a flight delay. For the Bureau of Transportation Statistics (BTS) (2016a) a flight delay is defined as a flight that is more than fifteen (15) minutes late than the scheduled time.

Airports and airlines are often associated with the image of a city or country, having a considerable impact on the local, national and international economy (Guimerà & Amaral, 2004). A study, developed by Ball et al. (2010), estimated the total impact of the costs of delays in the American economy by 32.9 billion dollars. This cost is composed of two types of costs, indirect and direct.

[2] Robert Toru Kiyosaki is an American author, public speaker and investor (for more info go to http://www.woopidoo.com/biography/robert-kiyosaki/).

Indirectly, the inefficiency in the airline industry increases the cost of doing business to other sectors, making the associated business less productive, reflecting a reduction of 4 billion dollars in the country's Gross Domestic Product (GDP). Directly, delays represent a cost of 28.9 billion dollars where:

- 16.7 billion dollars represents the passenger component (lost time, delayed flights, missed connections, among others);
- 8.3 billion dollars refers to the airline component (technical team, fuel, maintenance, among others);
- 3.9 billion dollars relates to customers who avoid traveling because of delays.

This phenomenon, in addition to quantitative costs, also has qualitative costs that influence the former one. For the passenger, affects his plans, which can cause displeasure regarding the company. According to the Aviation Consumer Protection Division (ACPD) from the Office of Aviation Enforcement and Proceedings (OAEP) (2016) between January and December of 2015, six thousand, four hundred and thirty-three (6433) consumer complaints were reported regarding flight problems such as cancellations, delays, and missed connections. As a result, representing a qualitative cost to the airline, demand, and reputation may be adversely affected when there is competition on the same route because passenger's choice of airline can be based on past events (Mazzeo, 2003).

The concept of traveling has been shaping over time. In the past, it was seen as a privilege. Over the years the circumstances have changed, and today, travel often represents a necessary evil. A result of air delays, increased security measures, and degradation of services provided (Ball et al., 2010).

The analysis of air delays becomes vital since a better knowledge of their existence, and corresponding triggers, can improve the performance of airlines and, consequently airports, in their operations by the possibility of anticipation (Yablonsky et al., 2014) and construction of schedules for example.

It should be considered the analysis of the delays with focus on the arrivals since these are more related with the passenger's satisfaction and because one arrival delay may trigger a delay in a departure (Tu, Ball, & Jank, 2008a).

Based on reports from several airports around the world, passenger traffic results for the busiest airports in 2015 put on top the Hartsfield-Jackson International Airport in Atlanta. Year after year, it showed a growth of 5.5% of passenger traffic reaching a record of more than 100 million passengers in that same year. This airport benefits from its strategic location being a major "gateway" to entry into North America, and also distance itself of a two hour flight from 80% of the population of the USA. In the 2015 International Airports Council report it occupies the first place both in the ranking of total passenger traffic – 101 491 106 passengers (Table 1) – and in the ranking of aircraft movements (Table 2) (Airports Council International, 2016).

Rank	*Airport city*	*Passengers*
1	Atlanta	101 491 106
2	Beijing	89 938 628
3	Dubai	78 010 265

4	Chicago	76 949 504
5	Tokyo	75 316 718

Table 1: Total passengers traffic in 2015
Source: Made by the author, adapted from (Airports Council International, 2016)

Rank	*Airport city*	*Aircraft Movements*
1	Atlanta	882 497
2	Chicago	875 136
3	Dallas	681 244
4	Los Angeles	655 564
5	Beijing	590 169

Table 2: Aircraft movements in 2015
Source: Made by the author, adapted from (Airports Council International, 2016)

1.2. OBJECTIVE

To characterize a flight, data as the information of airplane number, the airline company, the origin and destination, the schedule and actual departure and arrival time, the weather conditions, among others are typically used (Abdel-Aty, Lee, Bai, Li, & Michalak, 2007; AhmadBeygi et al., 2008; Belcastro et al., 2016; Choi et al., 2016; Ionescu, Gwiggner, & Kliewer, 2016; Khanmohammadi, Tutun, & Kucuk, 2016; M. S. Kim, 2016; Y. J. Kim et al., 2016; Klein, Craun, & Lee, 2010; Mueller & Chatterji, 2002; Pyrgiotis, Malone, & Odoni, 2013; Qianya, Lei, Rong, Bin, & Xinhong, 2015; Rebollo & Balakrishnan, 2014; Xu, Donohue, Laskey, & Chen, 2005; Yao, Jiandong, & Tao, 2010; Zonglei, Jiandong, & Tao, 2009). This data can become valuable when applied in a model for forecasting delays in future flights.

Thus, with this study, it is intended to predict the occurrence of a delay in the arrivals of Hartsfield-Jackson International Airport based on the referred variables, further ones considered in the Methodology chapter, and their respective contribution to the delay.

The steps required to reach the final objective undergo:

- Construct a database with information concerning the flights and additional information;
- Explore the delays accordingly to different variables;
- Construct a predictive model using Data Mining and Machine Learning techniques to predict the delay of a flight in the arrival;
- Apply the model developed in new data to make predictions and see which fits better to the problem according to the desired advance of the prediction;
- Identify the variables that contribute most to the existence of delay.

At the end of this project is expected to reach an algorithm that performs better in the data according to the desired advance of the prediction. Subsequently, these results can be compared to results presented in earlier studies with the same context of this work.

1.3. STUDY OUTLINE

This project presents an introductory chapter explaining the context in which the theme is inserted and nowadays relevance, as well as the main objectives adjacent to this work.

Chapter two presents a literature review about the need for KDD, Data Mining, and Machine Learning when dealing with issues where large volumes of data are inherent, such as airborne delays. It also presents state of the art, about the study of air delays, both at an industrial community level as well as at the scientific community level.

The third chapter exhibit the various stages defined for the development of this work. Here, the scope of the study is defined. The entire process of making the data available, processing and construction of the dataset is explained. Also, the various decisions made regarding the data preprocessing and transformation as well as the selection of the algorithm are described.

In the fourth chapter, the results of the application of the methodology chapter are illustrated and analyzed. The best approaches are selected for each of the algorithms chosen, and for each of the hypothesis of advance of the prediction. Furthermore, our results are compared to studies in the same area of research and with the same target type variables, but also to websites that predicts delays.

Following is the fifth chapter, where is present the conclusions underlying this work, as well as the chosen of the best model among all algorithms and approaches. Limitations and future works are proposed in the sixth chapter.

2. LITERATURE REVIEW

We are drowning in information but starved for knowledge.

– John Naisbitt[3]

According to Yablonsky et al. (2014) in June 2010 an inquiry was made about the twelve most troublesome aspects of travelling. The flight delays held the seventh place with a score of 6.8 according to the scale of 1 (less annoying) to 10 (more annoying).

Since 2010, the amount of delays has been fluctuating, with the year of 2014 having the highest percentage of delay. In 2015, the percentage of delayed flights was 19.53%, the third highest value since 2010 (Figure 1).

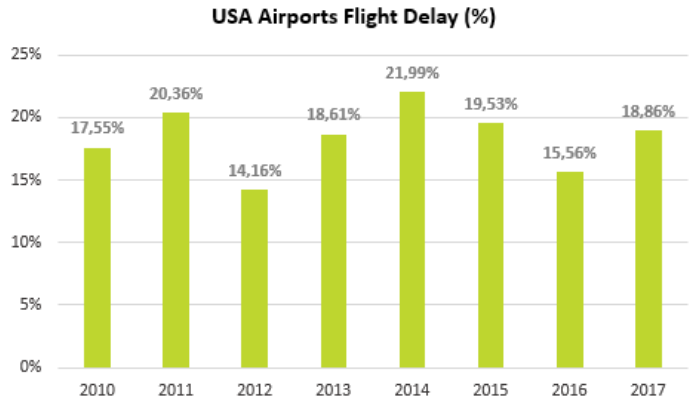


Figure 1: Percentage of total delayed flights per year in the USA, 2010-2017
Source: Made by the author, adapted from (Bureau of Transportation Statistics, 2017)

Over time it has become a phenomenon of great importance. As in other areas, there is a growing number of unstructured records that have been stored and do not add value to their pure state. Witten, Frank, & Hall (2011a) mentioned that we are overloaded with data and cannot control the exponential growth around us. More precisely, they mentioned that the amount of data in the world's databases doubles every twenty (20) months. In a business context, the large volume of data prevents the effective use of the same to create business value, competitiveness and efficiency. This made the lack of understanding in how to reach advantageous information a colossal obstacle (Lavalle, Lesser, Shockley, Hopkins, & Kruschwitz, 2011).

As a result, a gap is created between the production of data and our understanding of it. It is crucial to overcome this deficit because in data there is information beneficial to decision making, and consequently knowledge. To acquire knowledge from data the use of tools is required to allow the discovery of hidden information in these databases. According to Fayyad, Piatetsky-Shapiro, & Smyth (1996), the field of **Knowledge Discovery in Databases**, known as KDD (Knowledge Discovery Database) is the provider of the necessary tools and theories.

KDD consists of the process of discovering new, valid, useful and perceptible patterns in the data (Fayyad, n.d.; Mariscal, Marbán, & Fernández, 2010). Covers phases such as (1) the selection or

[3] John Naisbitt is an American author and public speaker in the area of futures studies (for more information go to https://en.wikipedia.org/wiki/John_Naisbitt)

creation of a dataset used as the basis for the discovery of standards. (2) The preprocessing of the same ones where unnecessary information is eliminated and corrected to assure the coherence of the same. (3) The transformation of data by reducing dimensionality or transforming existing ones. (4) The application of **Data Mining** (DM) where one looks for patterns in the data depending on the objective. And, finally, (5) the interpretation and evaluation of the results (Azevedo & Santos, 2008).

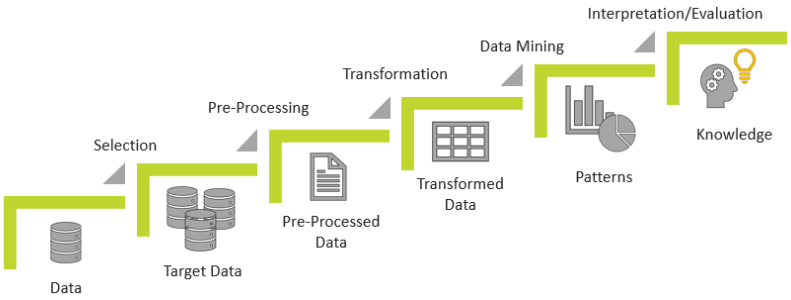


Figure 2: KDD Process
Source: Made by the author, adapted from (Kononenko & Kukar, 2007b)

Thus, DM is considered one of the steps of KDD as presented in (Fayyad, n.d.; Fayyad et al., 1996; Han & Kamber, 2011; Kononenko & Kukar, 2007b; X. W. X. Wang, 2009). It represents the application of algorithms to extract patterns from the data, translating in information (Fayyad et al., 1996). It has two main objectives. The first, forecasting, where future data are predicted about a particular variable of interest based on other variables present in the database (past information). And the second, description, where the focus goes through the discovery of patterns hidden in the data (X. W. X. Wang, 2009).

As a result, resort to DM and **Machine Learning** (ML) techniques have gained importance in solving these problems. The use of DM in this type of study is because it is a discipline that focuses on the discovery of knowledge in the data having as one of the objectives, the prediction of unknown data or future events as already mentioned (Kantardzic, 2011). On the other hand, ML is used for DM since it is a type of approach to the discovery of knowledge in the data. Focus on the construction of computational algorithms that can learn through data (from the past) to make predictions (Witten et al., 2011a). These disciplines, represented in Figure 3, enable better decisions and actions in real time without human intervention because of their high-value predictions (SAS, 2017). Belcastro, Marozzo, Talia, & Trunfio (2016) claim that the use of ML techniques associated with DM tools can help to perceive complex phenomena as well as solve various problems from many areas.

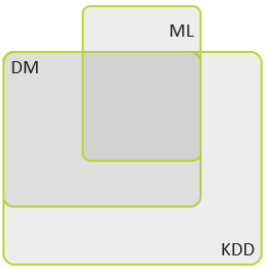


Figure 3: Relation between Knowledge Discovery Database (KDD), Data mining (DM) and Machine Learning (ML)
Source: Made By the author, adapted from (Kononenko & Kukar, 2007b)

The importance of flight delays made them the center of many investigations by both the **industrial** and **scientific** community.

An example of the first strand is Kaggle. Kaggle is a platform for analysis and predictive modeling competitions. There, companies and researchers publish data for people that have interest and knowledge in the field to compete in the production of models where can be rewarded with monetary awards.

This type of mechanism provided GE Aviation, in collaboration with Alaska Airlines, to launch a contest. The contest was the GE Flight Quest, where 173 teams competed with their algorithms to get a model that predicted delays with good performance (GE Aviation, 2012). DOT, as well as other companies, have provided data on air delays and cancellations so that users could predict which airlines were the best to travel (U.S. Department of Transportation, 2017).

On the other hand - scientific area -, several approaches were proposed by researchers to forecast and model delays. These approaches vary according to the different objectives of the forecast regarding the following three aspects:

- Network – when looking at the delay in several airports and the impact of it on the level of all airports, or a group of them;
- Airport – when the focus is on the study of the state of the delay, for example, about an airport - and, lastly;
- Flight – when is wanted to predict the delay of each flight.

It is also possible to distinguish the type of approach used in the development of these studies (Sternberg, Soares, Carvalho, & Ogasawara, 2017), such as:

- Statistical Analysis (SA) – covers the use of regression models, correlation analysis, econometric models, parametric and non-parametric tests, multivariate analysis, among others;
- Machine Learning (ML) – consists of a research/investigation that explores the development of computational algorithms that can learn from data and provide predictions from them;
- Operational Research (OR) – includes the development of advanced analysis models (e.g., optimization, simulation, queue theory, among others) to help stakeholders make better decisions;
- Probabilistic Models (PM) – covers analysis tools that estimate the probability of an event based on historical data.

For an understanding of what each author studied, and what type of approach implemented (it can be seen a summary in Figure 4) it is important to contextualize it.

In this way, at a **Network level**, several investigators used **Machine Learning** methods to understand the delays. Rebollo & Balakrishnan (2014) predicted delays in two ways. Through classification where they classified a departure delay as being more or less than a predefined threshold. And, through regression, where they estimated a future delay in departure on a specific route. Results of the study showed an average precision of 81% in the performance of Random Forests, algorithm selected for the classification, and a mean prediction error of the regression of 21 minutes. It should be noted that, for both means of forecasting, the error increased as the forecast advance also increased.

Xu, Donohue, Laskey, & Chen (2005) have created a methodology for representing and analyzing how system-wide effects arisen from subsystem-level causes through a Bayesian Network to estimate the delay propagation. The model created used an equation of linear regression as a priori probability (error rate of 19.1%) having superior performance about the other methods studied. The other methods comprise models where the Bayesian Network was estimated with the parameters of the

training data using a uniform prior distribution (38.1% error rate), and where the model was based on a linear regression (error rate of 61.9%). Thus, the authors concluded that the method could be extended by including more airports.

Qianya, Lei, Rong, Bin, & Xinhong (2015), and Rong, Qianya, Bo, Jing, & Dongdong (2015) have presented a method of analysis for flight delays also based on Bayesian Networks that could analyze and predict the delay during a flight. They analyzed the performance of the prior probabilities, created from (1) historical flight data statistics, (2) posterior probabilities using the Expectation-Maximization (EM) algorithm based on the finite Gaussian mixture model and (3) real data. They concluded that prior probabilities were more accurate (81.95%) proving that Bayesian Networks are an effective method for analyzing flights delays and that, as in the previous study (Xu et al., 2005), they have a great value to analyze system-level effects that result from lower-level causes.

AhmadBeygi, Cohn, Guan, & Belobaba (2008) analyzed the potential for delays to spread due to airlines. For that, they used propagation trees to compare two airlines. They concluded that the key points that had an impact on slowing the spread of delay were when cabin crews end their shift, and when cabin crew and airplanes were always together on all planned flights.

Zonglei, Jiandong, & Tao (2009) built a recommendation system to alert airports by monitoring related airports and informing the status of the delay.

However, other researchers have resorted to **Operational Research** methods to realize the delay. Pyrgiotis, Malone, & Odoni (2013) considered the propagation of delays in the airport network through a delay propagation algorithm (DPA). This tracked the spread of airport-calculated delays and their impact on the airport network. The authors mentioned that the goal was not to reproduce the exact delays but rather the trends and behaviors that are observed in the NAS system - network.

In Ionescu, Gwiggner, & Kliewer (2016) the main objective was to understand the potential of delay modeled through data used in the robust programming of airline resources. They provided a regression modeling approach for daily delay patterns based on the detection of spatiotemporal patterns in historical data. As a result, rules emerged, where precision was assessed through statistical modeling and compared to non-parametric random forests. They presented, given all decision rules as a whole, a 62.9% accuracy that could be compared to the random forests. However, the latter present an individual categorization and selection of rules as well as a lack of interpretation accordingly to the authors, contrary to the rules presented by them. They added that a possible delay might already have been taken into account through airline scheduling decisions, which is normal, questioning the generalization of their results to other data from other airlines.

Tu et al. (2008a) and Mueller & Chatterji (2002) used **Probabilistic Models** to model their problems. The former developed a strategic model to estimate delays in departures. They used non-parametric methods for daily and seasonal trends of delay propagation to predict departure delays, and mixture distribution to estimate residual errors, used to calculate the probability of the delay. According to the authors, it presented a reasonable adjustment quality, robustness in the choice of parameters and good forecasting capabilities and could be easily adapted to other airport /airline combinations.

The latter set out to improve the accuracy of forecasting delays by calculating the probability of departure, en-route and arrival delay using Poisson and Normal distributions. This study resulted in

several delay metrics for the analysis of an airport network (all ten airports experiencing significant delays), based on individual airports through its 21-day review period. They concluded that the departure delay is better modeled through a Poisson distribution whereas, the en-route and the arrival delay fit the Normal distribution better. In addition, the authors also used the **Probabilistic Models** for analysing the delay at a **Airport Level**.

As previously mentioned, Pyrgiotis, Malone, & Odoni (2013) used **Operational Research** to analyze the spread of airport delay and network impact. They used the Approximate Network Delays (AND) model composed of queueing theory (QE) to calculate the individual airport mode delay and to be able to analyze the network.

Yablonsky et al. (2014) and Klein, Craun, & Lee (2010) resorted to **Statistical Analysis**. The first ones focused on the average delay time at Hartsfield-Jackson Atlanta International Airport based on data for several years. The main objective was to determine the annual cyclical delays. They considered that the information resulting from their study was relevant when forecasting periods of air delays. Concluded that there is an annual pattern of delays being caused by the volume of flights at the airport and the amount, and frequency, of precipitation occurring in Atlanta.

The second ones, developed a model through multiple linear regression focusing in climate-related delays. They were based on a metric that helps to estimate the impact of climate on flight schedules, and an impact metric that helps predict expected weather in the flight time, WITI (Weather Impacted Traffic Index) and WITI-FA (Weather Impacted Traffic Index - Forecast Accuracy), respectively. They claimed that 70% of the delay could be explained by the WITI factors used as explanatory variables and that the model predicted the time and magnitude of the climate impact on delay successfully.

On the other hand, there were also those who used **Machine Learning** to predict the delay at the airport level. Zonglei, Jiandong, & Tao (2009) focus on predicting the seriousness of the delay at specific airports to monitor airport delays and alerting specific airports through a recommendation system. In this case a Chinese airport was analysed to understand how this delay could influence the delay at a network level and other airports related to it. To predict the seriousness of the delay at China's airport, they based on the K-Nearest Neighbor algorithm using historical data to compare and recognize similar situations in the past. The authors concluded that, since the forecast is based on the comparison and analysis of historical data, the proposed model reflected a precise and rapid forecast offering logical explanations. Previously, Zonglei, Jiandong, & Guansheng (2008) already had presented a method, where instead of based on a recommendation system they based on machine learning, however, it also served as a large alert for flight delays. They used non-supervised learning methods, more specifically clustering with the K-means algorithm, to extract classes of airport delay, and then used it to predict the class of delay of the airport in each day by applying a supervised learning method to the data. They compared several supervised learning methods to predict the delay class, and decision trees were the ones that stood out with 80% of confidence.

Smith & Sherry (2008) developed a process using the Support Vector Machine (SVM). With it a climate forecast for a particular area could be used as input in estimating an airport's delay. Having also the ability to predict the impact of weather on future flights, i.e. how long could expect a flight to be delayed due to the weather. This model, according to the authors, showed to be correct 83% of the time.

Yao, Jiandong, & Tao (2010) have focused on predicting the delay propagation - through a proposed algorithm - caused by aircraft, cockpit and cabin crew. The predicition was for situations where it is necessary to wait for these elements. For example, if a flight is delayed and, at its destination, the aircraft or cabin crew are needed for another flight, they will delay the latter as a domino effect. Used the results to create and provide an alarm rank of flight delays to airports. They stated that their model and algorithm could be used to efficiently calculate the propagation of delays caused by required flight resources on the same flight.

Balakrishna, Ganesan, Sherry, & Levy (2008) applied a Reinforcement Learning (RL) algorithm to estimate the taxi-out time (time between the departure of a plane from the boarding gate and the time it takes flight). The Markov decision process was used to model the problem being solved through Machine Learning Reinforcement Learning algorithm. The authors were able to achieve 60% accuracy of the model.

Y. J. Kim, Choi, Briceno, & Mavris (2016) have proposed recurrent neural networks as a method of predicting the status of day-to-day airport-level delay due to capturing sequential and temporal relationships in the data. Stated that the application of Long Short-Term Memory Recurrent Neural Networks (LSTM RNN) architectures in the forecast model allowed a more reliable acquire on one-day delay status of an aiport.

However, the area of interest in this study is the prediction of delay in an **Individual Flight level** and not the ones mentioned above. As such, some studies have been proposed in which the delay of an individual flight was the objective to the forecast.

Some authors used **Statistical Analysis** to understand and detect patterns on a flight. Abdel-Aty, Lee, Bai, Li, & Michalak (2007) have evaluated the performance on the arrival of a flight, through a two-stage approach using mathematical frequency analysis. They were able to identify patterns of delay where it was possible to determine which were the most important variables that affected the delay. Subsequently, the relationship between the variables and the delay was investigated through statistical techniques in which the periodic patterns of delays were examined. Through their results, they observed that the delay of a flight was associated with the precipitation, flight distance, time of the year, the day of the week, time of arrival and space of time between the arrival of two successive flights. Accordingly to the authors, their model could be adjusted itself to any type of data.

M. S. Kim (2016), concluded that the Spline Smoothing regression model surpasses the linear regression and median regression models in prediction performance. It adjusted better to the data and was the one that represented the delays both in the long term and in the short term. He considered that the variables of delay departure, flight time, airline, weather conditions and time of the year were relevant in the accuracy of the forecast, and the use of the delay variable at departure significantly improved the accuracy. Later, included in his study the area of **Probabilistic Models**, and suggested a method to calculate the probability of the time of arrival of a flight adjusting the residuals of the model to a distribution Skew t.

Other authors resorted to **Machine Learning** for the purpose being the focus of our study.

Y. J. Kim, Choi, Briceno, & Mavris (2016) as already mentioned, in addition to predicting the status of the day-to-day delay at an airport level used neural networks to be able to predict the class of the

delay of an individual flight. Gathered data about the flight data (day, season, month, date, origin, destination, schedule times) and weather data (wind direction, wind speed, cloud height, visibility, precipitation, snow accumulation, intensity, descriptor and observation code). Furthermore, allied the status of the day delay computed in the first stage of their study. They showed that their model achieved 87.42% accuracy, better than the best predictions until then demonstrated, 83.4% and 81% (Choi, Kim, Briceno, & Mavris (2016) and Rebollo & Balakrishnan (2014), respectively). Due to the acquisition of a more reliable day-delay status using Long Short-Term Memory Recurrent Neural Networks (LSTM RNN) architectures, the forecast model of the delay state of an individual flight also became more accurate.

Choi, Kim, Briceno, & Mavris (2016) also presented a classification model in which the main objective was to exclusively predict flight delays of individual flights caused by climatic conditions. They used data collected from the BTS airline On-time Performance dataset for the years of 2005 to 2015 using features like flight schedules and day. Also included weather conditions variables obtained by the Integrated Surface database of the National Oceanic and Atmospheric Administration (NOAA), at the origin (Denver International Airport) and the destination (Charlotte Douglas International Airport). For the purpose, resorted to the use of algorithms of data mining and machine learning. Considered that the best classifier was the Random Forest based on their results. They also studied the accuracy of the results for different horizons (5, 1 and 0 days) and concluded that results with real climatic conditions have better accuracy (26.79%, 30.36%, and 80.36% respectively). They mentioned that accuracy is higher when sampling technique is not applied and understand that this happened because classifiers are biased towards the on-time class, the majority.

In the same line of thought, Belcastro, Marozzo, Talia, & Trunfio (2016) applied a parallel version of the random forest algorithm to predict the delay in the arrival of a flight due to the weather. The main objective was to be able to predict, with a few days in advance, the delay in the arrival of an individual flight due to the weather. They used data about the flight information such as schedule times but also origin and destination from the airline on-time performance (AOTP) dataset from RITA-BTS comprising the years of 2009 till 2013. Joined variables of weather conditions (temperature, humidity, wind direction and speed, barometric pressure, sky conditions, visibility and weather phenomena descriptor) at the origin and destination accordingly to the flight timetable acquired from the Quality Controlled Local Climatological Data (QCLCD) dataset from the National Climatic Data Center. Finally, they stated that the model had an accuracy of 85.8% for a threshold of 60 minutes and that even if they did not consider the weather the model would achieve an accuracy of 69.1%. Nonetheless, when considering a threshold of 15 minutes, it achieved 74.20% of accuracy.

The work presented here, in contrast to others already mentioned, intends to develop an algorithm to predict the delay in an individual flight taking into account (1) flight information, similar to Y. J. Kim, Choi, Briceno, & Mavris (2016) and Zonglei, Jiandong, & Guansheng (2008); (2) climate at origin and destination, similar to Belcastro, Marozzo, Talia, & Trunfio (2016) and Choi, Kim, Briceno, & Mavris (2016), but also; (3) information of the aircraft as well as (4) possible congestion of the system. Having the purpose of understanding if it is possible, to improve the performance of the models already presented in this type of approach (machine learning) and with this objective (individual flight prediction).

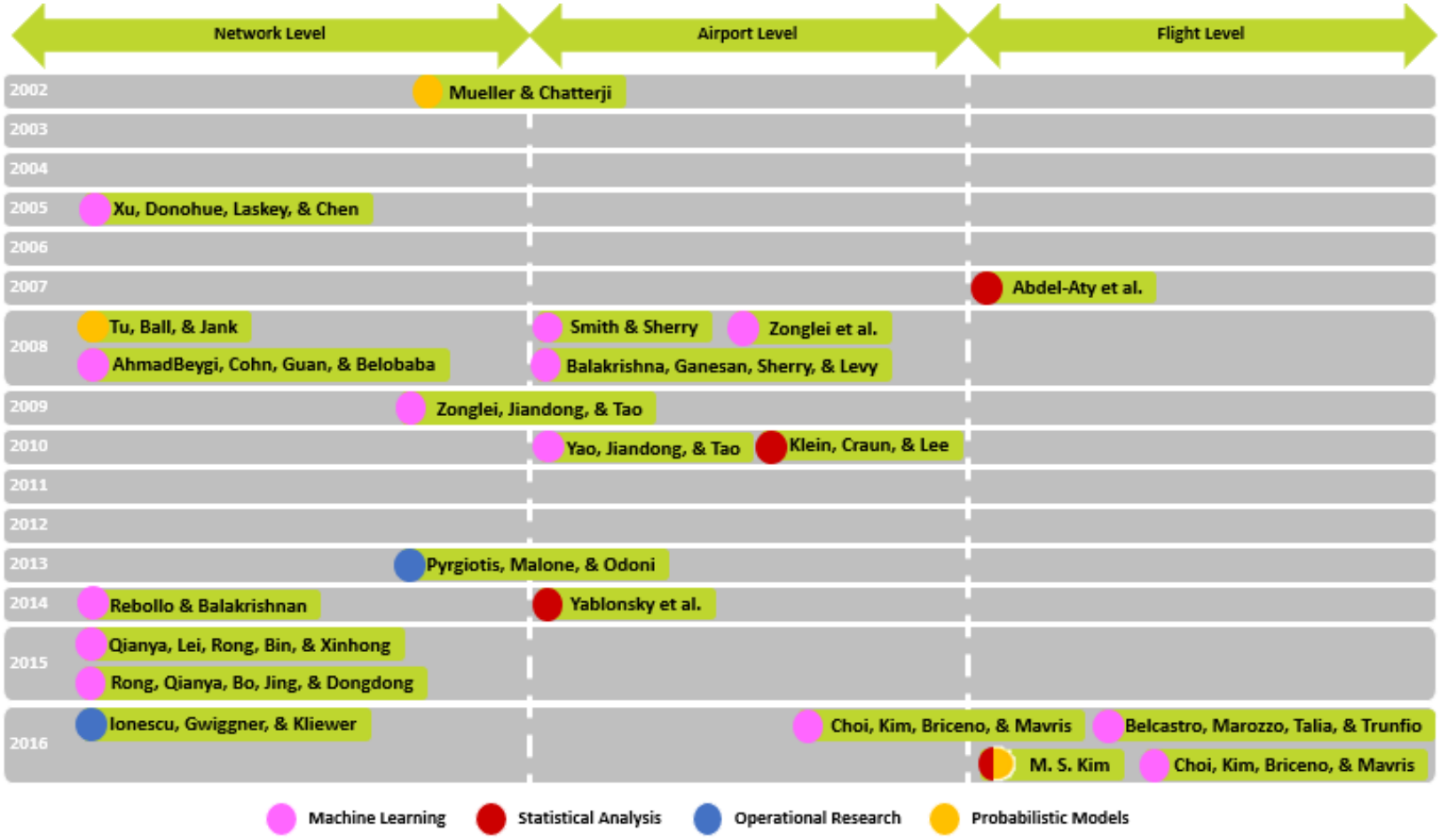


Figure 4: Literature Review
Source: Made by the author, adapted from (Sternberg et al., 2017)

3. METHODOLOGY

Information is not Knowledge.

– Albert Einstein (1879-1955)[4]

This section presents a methodology that describes the data sources, the validation process of input data formats and following treatment for integration in the construction of the final dataset. Subsequently, a description of the methods used for partitioning the dataset as well as the type of sampling is given. It is also presented the procedure for pre-processing and transforming the variables. And finally, an explanation of the classification methods that will be applied, as well as the techniques of evaluation of the classifiers used to assess their performance is presented as illustrated below in Figure 5.

Figure 5: Work Project Methodology
Source: Made by the author

For the development of this project, different tools for each phase were needed. The tools for the collection and integration of the data – first phase – were Microsoft Excel and SQL Server 2014 Management Studio.

For the second phase, Waikato Environment for Knowledge Analysis software, also known as Weka, was used. It is an open source software developed at the University of Waikato in New Zealand, and it is considered a significative collection of machine learning algorithms and data preprocessing tools. It supports KDD and DM processes through an interface where users can compare different methods and identify those that best fit the problem under analysis (Witten, Frank, & Hall, 2011b)(Table 3).

For Zupan and Demsar (2008) it is a well-known open source software for ML and DM. It allows for advanced users to use it through Java programming or its programming interface, Command-line interface (CLI). And for other users, with fewer skills of programming, it can be used through its graphical interface, KnowledgeFlow, which provides the design of the flow for data loading and processing, application of algorithms, their evaluation and visualization of results. Also through its slightly more limited but easy-to-use interface, where preliminary data experience is enabled, Explorer. And, its interface where is easier to compare the performance of different classifiers at the same time and identify which is the most appropriate for the problem, Experimenter (M. Hall et al., 2009; Witten et al., 2011b). It has become a reference and a growing tool for the ML community attracting many users and researchers (Muenchen, 2017).

[4] Albert Einstein was a German theoretical physicist (for more information go to https://en.wikipedia.org/wiki/Albert_Einstein)

KDD	*Weka*
Pre-KDD	-
Selection	Data sources
Pre-processing	Preprocessing tools/Filters
Transformation	
Data Mining	Learning Algorithms
Interpretation/ Evaluation	Evaluation methods/ Visualization modules
Post KDD	-

Table 3: Implementation of KDD in Weka
Source: Made by the author

3.1. SELECTION

In this sub-theme, the necessary steps to get to the final dataset, containing only the data of interest for the study, are covered. First, the scope of the study and the time interval defined in the extraction of the data are mentioned. Subsequently, the presentation of the data is done by describing the sources from which the raw data, necessary to assemble the final dataset, were extracted. The validation of data consistency, due to the need of merging data from different sources. And at last, the rationale followed for the integration of the three sources in a single database as well as the construction of the final dataset with the presentation of the variables chosen and their description are presented.

3.1.1. Study Scope

3.1.1.1. Geographical

The geographical scope of the study is the United States of America. The flights that were used as study object were the domestic arrival flights to Hartsfield-Jackson Airport, in the city of Atlanta, which is part of the State of Georgia located in the Southeast area of the USA.

3.1.1.2. Temporal

To guarantee the representativeness of all the seasons of the year and a good performance of the model, two months of each year of 2015 will be used as a study base, due to be the last full year at the time of data extraction. Therefore, based on Table 4, the months that were used were April, May, July, August, October, November, January and February.

Season of the year	*Start date*	*End date*
Spring	21st of March	20th of June
Summer	21st of June	20th of September
Autumn	21st of September	20th of December
Winter	21st of December	20th of March

Table 4: Seasons of the Year in the Northern Hemisphere
Source: Made by the author

3.1.2. Data

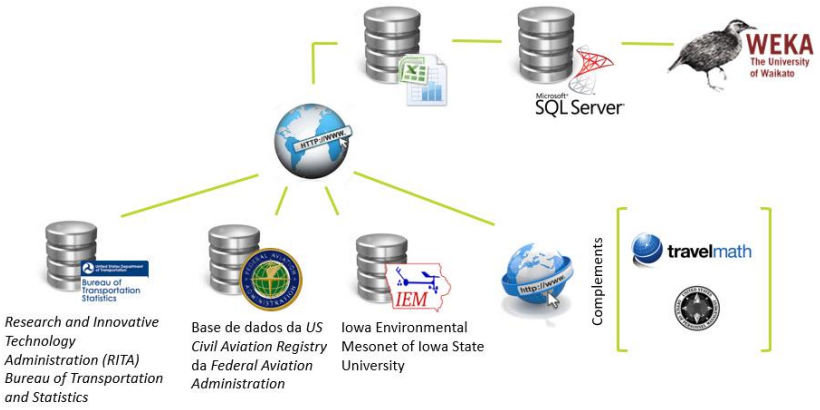


Figure 6: Data acquisition and preparation steps for upload in WEKA
Source: Made by the author

3.1.2.1. Sources

Since the centrality of information in a single source is not possible due to the need of different types of data that could represent the possible factors causing the delays, the collection of the data from different platforms described below was carried out.

A. <u>Bureau of Transportation Statistics (BTS)</u>

The Bureau of Transportation Statistics provides compiled data on U.S. transportation systems. It also improves the quality and effectiveness of the Department of Transportation's statistical programs through research, development, and promotion of improvements in data acquisition and use. Is part of the Research and Innovative Technology Administration (RITA) - a unit of the US Department of Transportation (DOT) - and is one of the lead agencies in the U.S. Federal Statistical System. The data of interest comes from service of the BTS, Transtats, a database with information on on-time performance. From this source information about flights can be extracted (Bureau of Transportation Statistics, 2016b);

B. <u>Federal Aviation Administration (FAA)</u>

The U.S. government is responsible for regulations and all matters related to civil aviation in the country from the construction and operation of airports to the air traffic management and certification of personnel and airplanes, among other responsibilities. It provides different databases: the one used to obtain information about the airplane used in a particular flight was the one referring to the U.S. civil aviation register (Federal Aviation Administration, 2016a);

C. <u>Iowa Environmental Mesonet (IEM)</u>

Iowa Environmental Mesonet is a website of the Agronomy Department of Iowa State University of Science and Technology that aims to gather, collect (from existing resources), compare, disseminate and archive observations (Department of Agronomy, 2017a). It is a valuable resource for anyone who searches for overlaid information as well as for historical comparisons. It puts data in useable formats which others can use it and share being the number one provider of weather service information to the National Weather Service (Herzmann, Klein, & Taylor, 2013). To acquire climatic data, the archive of automated airport weather observations from the entire world maintained by

the IEM was accessed. These observations are called ASOS (or METAR - Meteorological Terminal Air Report -, a generic term format for reporting weather information), provided by the Automated Surface Observing System (Department of Agronomy, 2017b). This type of system is located at airports for aviation support and weather prediction, providing essential observations for the National Weather Service (NWS), the Federal Aviation Administration (FAA), and the Department of Defense (DOD) (Department of Agronomy, 2017c; Nadolski, 1998).

D. Complements

In order to acquire additional data, other two different sources to complement the three initial ones for the creation of further variables were used.

- Travel Math used to attain information on the time zone between origin-destination pairs (TravelMath, 2017a) and also to obtain information about the time flight duration of a route (TravelMath, 2017b);
- Office of Personnel Management (OPM), that "provides human resources, leadership, and support to Federal agencies and helps the Federal workforce achieve their aspirations as they serve the American people" (U.S. Office of Personnel Management, 2017a). This was consulted to access the federal holidays on the year of 2015 (U.S. Office of Personnel Management, 2017b).

3.1.2.2. Data Validation and Limitations

The use of data from different sources can lead to inconsistency and, consequently, lack of compatibility. From this result the Merge/Purge problem, a term designated by the companies, and explained as the existence of data fields that may be different between datasets or that may be incorrect (due to representation, unity, along with others)(Hernández & Stolfo, 1998).

For this phase and to avoid the problem above mentioned, a significant amount of time was spent preparing and analyzing the veracity and consistency of the vast volume of data acquired from the three different sources, making their compatibility possible.

Regarding the flight information, there was a need to provide understanding, and to allow further calculations between variables for the creation of new ones. This was accomplished by modifying the way variables were represented turning them into specific variables of time. This also made easy querying between tables for further variables creation possible.

When dealing with the need for information about the airplane, it was necessary to validate the key variable between the RITA database and the information needed: The N-Number. It refers to a single alphanumeric number registered within the National Aviation Authority (NAA) which identifies the airplane and the country in which it was registered (Federal Aviation Administration, 2015).

Regarding the weather information, duplicate records elimination was necessary as well as the correction of information of present weather codes assumed by the Microsoft Excel as functions and not text.

The necessary transformations were possible using Excel formulas and its programming language VBA. More detailed information can be consulted in Annex 1 on the Annexes chapter.

3.1.2.3. Dataset Construction and Description

Given (1) the need to integrate all data described above; (2) the possibility of crossing it on the same place; (3) the wish for searching and selection of specific data, and (4) due to the high number of records collected, it was necessary to create a database first.

In order create the database, a design of a relational entity model was done. This model was initially proposed by Chen (1976), due to its enormous use and acceptance in the area (Tryfona, Busborg, & Christiansen, 1999) and to the basis it provides for a single view of the data (Chen, 1976). The choice of this model was based on its main characteristics: simplicity and clarity of concepts (Fahrner & Vossen, 1995). It represented information, through its concepts, regarding entities, attributes and associations (relations) (Teorey, Yang, & Fry, 1986).

The steps required for the design of the database model consist of (1) understanding what is necessary to build it, regarding entities and relations of interest (e.g., identifying all the entities that are needed, among all the information available, to construct the final table). (2) Identifying information that can classify the type of relationships between entities (e.g., understandiing the relationships between information of weather and flight observation locations as well as between routes and flight information). (3) Defining the attributes of each entity (e.g., for the weather entity the attributes were the variables extracted from the IEM website; the same happens with the flight entity that was composed by the attributes extracted from the RITA database). And, finally, (4) organizing the data into entity/relation relations and seeing the primary keys of each entity (e.g., understanding from which attributes we are able to link different entities to build the final table)(Chen, 1976).

Following the steps mentioned above, the physical model of the relational entity was designed, using the Crow's Foot notation. Through this notation, the primary and foreign keys, and the names of the tables and the columns were defined, as well as the type of data accepted in each column to understand exactly how the model would be created in the database and to have a defined idea of what would be necessary (Annex 2 of Annexes chapter).

From this model, it was necessary to resort to SQL Server 2014 Management Studio to implement the drawing by creating tables and their relationships. Subsequently, due to the non-existence of a dataset that responded to the existing need in this study, it was necessary to create one using the querying logic of the database to obtain a final table as a result of the query and intersection of existing tables in the database created.

Previously, in chapter one - Introduction -, some of the factors that caused the delays were presented. Based on these, the final dataset and the choice of its variables was made having in mind the importance and need of each one as the cause of delay in the flights.

For that reason, the independent variables that were chosen based on previous articles are presented below - distinguished by importance and purpose.

A. <u>Variables of flight information</u>

These variables were used to give details about each flight and situate it in a chronological space, making it possible to understand patterns in delay behavior by the time of the year, time of the day and by distances. All chronological variables, after being used to obtain other variables, were

adapted from hours and minutes to total minutes as it is applied by Kim (2016) to have a standard measure for all variables of time. With those goals, nine variables were established such as month, day of the month, weekday, scheduled flight duration, distance, scheduled departure time, real departure time, departure delay and scheduled arrival time.

B. Variables of airport information

These variables were used to give geographical information of each flight. For that purpose, only one variable was needed, the origin airport.

C. Variables of flight and plane information

The variables mentioned here were used as a way of identifying the airplane, the possibility of mechanical problems and information of possible volume of passengers as a justification of probable delays. In that order, six variables were provided such as the airline company, the flight number, the antiquity, the manufacturer, the model, and the maximum seats of an airplane.

D. Variables of flight delay propagation information

These variables were used as a means of perceiving whether the delay has already happen or could have happen at a time different from the departure time at the airport of origin which generates large volumes of passengers consequently. With this intent, two variables were created: the previous day delay occurrence in the origin airport, that accordingly to Rebollo & Balakrishnan (2014) affects the ability to recover from it when there is a large volume of delays and might result in higher air traffic values reuslting in departure delays that translate into arrival delays by consequence; And, the holiday occurrence that can cause a more significant influx of passengers causing a possible delays.

E. Variables of weather information

The variables created for this category were used as a means of informing the weather conditions in three situations (for a more detailed explanation consult Annex 6 of the Annexes chapter):

- At the origin, at the scheduled time of departure;
- At the destination, on the scheduled time of departure in origin, with the time zone of the destination;
- At the destination, at the scheduled time of arrival.

Variables with En-Route climate information were not considered as it is challenging to obtain climate information for all positions of the airplane due to the combination of the measurements of the reports of different stations. They were also not considered due to the need to take into account the altitude at which the plane is at (Takacs, 2014). In agreement with Mueller & Chatterji (2002), the weather is the largest contributor to delays in the air traffic control system.

These three phases are differentiated because the flight can be delayed and/or canceled by the weather. As such, before a flight departs, an assessment is made as to whether the conditions for taking off are met or not.

Three possibilities can make the flight not to take off or, after taking off, not land at the destination at the scheduled time, causing a delay in departure and arrival. Such possibilities are (1) the weather is not favorable at the scheduled time of departure at the origin. (2) The weather conditions are not favorable at the destination at the same time that the airplane is scheduled to depart at the origin

(and therefore is influenced by the time zone). And, at last, (3) The weather is not favorable at the destination at the time the plane is scheduled to land (Krozel, Capozzi, Andre, & Smith, 2003).

Additionally, and according to the website of the Bureau of Transportation Statistics (2016c), the total weather has a large share in the percentage of delayed flights (32.8% in 2015). The total share of the percentage of delays due to the weather for the BTS is the percentage of delayed or canceled flights due to two factors. First, the extreme weather category, which refers to cases that entirely prevent a plane from taking off (5% in 2015). And, secondly, the category of delays and cancellations assigned to the NAS, which include a time subcategory, in which case only those cases where the system can be delayed, but does not prevent take-off, are taking into account (more than half of the 22.9% in the year of 2015).

Takacs (2014), in his study, mentioned that precipitation, visibility, wind speed, and temperature were the most crucial weather features. For such reasons, there were created fifteen variables to portray the weather characteristics at the three phases for each weather condition: temperature, precipitation, wind, visibility, and event.

Table 5 summarizes all the topics, above discussed, where all variables defined are illustrated and their importance corroborated by other researchers.

It should be noted that not all information given about variables is complete and some articles do not explicitly state all the variables used. Thus, the percentage of variables used in this article, in relation to each of the articles below, is not always precise. For that reason, in some situations the percentage is preceded by an inferior signal ("<") representing that, in that article, the number mentioned is the percentage for the variables explicitly mentioned although, it is referred that other variables exist. In that way, as the number of variables increases, the percentage of variables used in this study decreases.

Variables / *Articles*	(Khanmohammadi et al., 2016)	(Mueller & Chatterji, 2002)	(Pyrgiotis et al., 2013)	(Rebollo & Balakrishnan, 2014)	(Xu et al., 2005)	(Yao et al., 2010)	(Klein et al., 2010)	(AhmadBeygi et al., 2008)	(Kim et al., 2016)	(Belcastro et al., 2016)	(Abdel-Aty et al., 2007)	(Ionescu et al., 2016)	(M. S. Kim, 2016)	(Qianya et al., 2015)	(Zonglei et al., 2009)	(Choi et al., 2016)
Month				x			x		x			x	x		x	x
Day	x	x					x		x				x		x	x
Weekday	x			x					x		x	x	x			x
Schedule Flight Duration		x				x							x			
Distance											x					
Schedule Departure Time	x	x	x	x		x	x	x	x	x	x	x				x
Real Departure Time	x		x	x			x			x	x	x		x		
Departure Delay	x				x						x		x	x		
Schedule Arrival Time	x	x	x	x		x	x	x	x	x	x	x				x
Origin Airport		x	x	x		x		x	x	x	x	x		x	x	x

Airline Company				x									x	x		
Flight Number				x	x		x				x					x
Antiquity																
Manufacturer																
Model														x		
Maximum seats																
Previous Day Delay Occurrence				x												
Holiday Occurrence																
Temperature at Origin										x						x
Temperature on Destination at Schedule Departure Time																
Temperature at Destination										x						x
Precipitation at origin							x		x							x
Precipitation on destination at Schedule departure time																
Precipitation at Destination											x					x
Wind at Origin							x		x	x						x
Wind on Destination at Schedule Departure Time																
Wind at Destination										x	x					x
Visibility at Origin							x		x	x						x
Visibility on Destination at Schedule Departure Time																
Visibility at Destination										x						
Event at Origin							x		x	x						
Event on Destination at Schedule Departure Time																
Event at Destination										x						
Used Variable Percentage (%)	46	56	67	<26	<17	17	<56	50	53	75	63	<86	67	56	<100	<78

Table 5: Synthesis of chosen variables and their use in other articles
Source: Made by the author

For DOT and, consequently, for BTS, as it for this study, a flight is considered to be delayed if the actual time of arrival is greater than 15 minutes in relation to the scheduled time of arrival. However, there is no universal definition of how it is measured.

In that sense, in the model of this study, as a dependent variable, the discrete binary variable of arrival delay is considered as having been used in other studies as seen in Table 6. It represents the existence, or not, of the arrival delay. If the actual time of arrival is greater than 15 minutes from the scheduled time then the variable assumes the value 1, otherwise it assumes the value of 0, according to the DOT definition.

Type of Dependent Variable	*Articles* (Khanmohammadi et al., 2016)	(Pyrgiotis et al., 2013)	(Rebollo & Balakrishnan, 2014)	(Smith & Sherry, 2008)	(Xu et al., 2005)	(Yao et al., 2010)	(Klein et al., 2010)	(AhmadBeygi et al., 2008)	(Kim et al., 2016)	(Balakrishna et al., 2008)	(Belcastro et al., 2016)	(Abdel-Aty et al., 2007)	(Ionescu et al., 2016)	(Qianya et al., 2015)	(Zonglei et al., 2008)	(Zonglei et al., 2009)	(Choi et al., 2016)
Continuous	x	x	x	x			x	x		x			x				
Discrete			B		C	C			C		C	B		C	C	C	B

Table 6: Synthesis of the type of dependent variables used in other articles
where B stands for Binary Discrete type of dependent variable and, C for Categorical Discrete type of dependent variable.
Source: Made by the author

Since this work studies the air delays, canceled or diverted flights are not considered since they do not represent information about the delay (Abdel-Aty et al., 2007; Belcastro et al., 2016).

Thus, the final input dataset is then composed of thirty-four (34) dependent variables to explain the independent variable. It has a total of 248 956 records for eight months in the year of 2015 (January, February, April, May, July, August, October, November) as illustrated in Table 7.

Variable	*Type*	*Observation Values*					*Role*	*Description*
		Min-Max	*N. levels*	*Mean*	*Mode*	*Missing Values*		
MONTH	Nominal	-	8	-	8	0	independent	Month relative to the flight
DAY	Nominal	-	31	-	13	0	independent	Day relative to the flight
WEEKDAY	Nominal	-	7	-	5	0	independent	Weekday relative to the flight (1-Monday; 7-Sunday)
SCHED_FLIGHT_DUR	Numeric	24 - 489	-	91,257	-	0	independent	Schedule flight duration about the origin-destination route (in minutes)
DIST	Numeric	83 - 4502	-	639,589	-	0	independent	Distance between origin and destination airports (in miles)
SCHED_DEP_TIME	Numeric	15 - 1439	-	738,353	-	0	independent	Schedule departure time (in minutes and in local hour)
REAL_DEP_TIME [5]	Numeric	0 - 1439	-	743,027	-	0	independent	Real departure time (in minutes and in local hour)

DEP_DELAY [5]	Numeric	0 - 1289	-	10	-	0	independent	Difference in minutes between Schedule and real departure time. Departures earlier to the schedule are represented as 0.
SCHED_ARR _TIME	Numeric	1 - 1439	-	869,689	-	0	independent	Schedule arrival time (in minutes and in local hour)
ARR_DELAY	Nominal – Binary (0,1)	-	2	-	0	0	dependent	Indicator of arrival delay. If the delay is superior to 15 minutes is represented as 1 otherwise is represented as 0.
ORIGIN	Nominal	-	168	-	MCO	0	independent	3 letters code of the origin airport, accordingly to IATA.
AIRLINE_COMP	Nominal	-	11	-	DL	0	independent	Airline Company unique code
FLIGHT_NUM	Nominal	-	3265	-	N844AS	0	independent	Flight N-number attributed by *National Aviation Authority* (NAA)
ANT	Numeric	0 - 56	-	16,244	-	8873 (4%)	independent	Airplane antiquity
MAN	Nominal	-	26	-	BOEING	1405 (1%)	independent	Airplane Manufacturer name
MOD	Nominal	-	106	-	MD-88	1405 (1%)	independent	Airplane model name
MAX_SEATS	Numeric	2 - 451	-	139,42	-	1464 (1%)	independent	Airplane maximum seat number
PREV_DAY _DELAY_OCURR	Numeric	0 - 810	-	34,9	-	1021 (0%)	independent	Number of flights canceled and delayed (departure) on the previous day at the airport of origin
HOLIDAY _OCURR	Nominal -Binary (0,1)	-	2	-	0	0	independent	Indicator of holiday occurrence. Represented as 1 if the same day or in a buffer of 3 days occurs some federal holiday, otherwise 0
TEMP_ORIGIN	Numeric	-49,4 – 45,6	-	16,217	-	940 (0%)	independent	Measured in Celsius degrees
TEMP_DEST _SCHED_DEP	Numeric	-11,7 - 35	-	17,165	-	237 (0%)	independent	

[5] Variable Dep_Delay and Real_Dep_Time used as a basis for the construction of two types of dataset one contemplating it and another without it because of its known importance on determining a delay in arrival as proved in (M. S. Kim, 2016) and for not allowing advance in the prediction.

_TIME								
TEMP_DEST	Numeric	-11,7 - 35	-	17,906	-	204 (0%)	independent	
PRECIP_ORIGIN	Numeric	0 – 71,37	-	0,053	-	852 (0%)	independent	Measured in millimeters per hour
PRECIP_DEST _SCHED_DEP _TIME	Numeric	0 – 24,89	-	0,08	-	237 (0%)	independent	
PRECIP_DEST	Numeric	0 – 24,89	-	0,08	-	200 (0%)	independent	
WIND_ORIGIN	Numeric	0 – 166,9	-	7,733	-	1335 (1%)	independent	Measured in miles per hour
WIND_DEST _SCHED_DEP _TIME	Numeric	0 – 32,2	-	7,85	-	433 (0%)	independent	
WIND_DEST	Numeric	0 – 32,2	-	8,13	-	386 (0%)	independent	
VISIB_ORIGIN	Numeric	0 - 99	-	9,239	-	1005 (0%)	independent	Measured in miles
VISIB_DEST _SCHED_DEP _TIME	Numeric	0,12 - 10	-	8,921	-	237 (0%)	independent	
VISIB_DEST	Numeric	0,12 - 10	-	8,957	-	200 (0%)	independent	
EVENT_ORIGIN	Nominal	-	162	-	BR	217477 (87%)	independent	Situation descriptor (METAR codes)
EVENT_DEST _SCHED_DEP _TIME	Nominal	-	33	-	BR	210125 (84%)	independent	
EVENT_DEST	Nominal	-	33	-	BR	210150 (84%)	independent	

Table 7: Overview of used variables
Source: Made by the author

3.1.3. Sampling Techniques

Most datasets are not made up of a similar number of observations from each class. When, in a classification problem, the classes are not represented equally we are facing an **unbalanced dataset**. This characteristic may reflect a very high accuracy of the model. However, having many observations of a class will cause the model to learn from the data, and always decide by the majority class resulting in a false classification accuracy (Figure 7) (Chawla, Bowyer, Hall, & Kegelmeyer, 2002; Hoens & Chawla, 2013; Lachheta & Bawa, 2016; Weiss, 2004).

One way to solve this is through sampling techniques where the goal is to create a dataset that has a similar distribution of classes so that the classifiers can correctly capture the division between the majority and minority classes (Hoens & Chawla, 2013). There are two types of methods for this, (1) copying

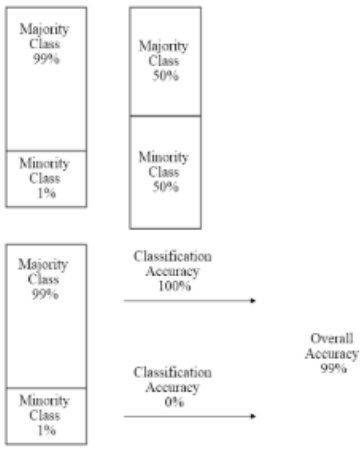


Figure 7: Problem of unbalanced classes
Source: (Lachheta & Bawa, 2016)

minority class observations - **oversampling** -, or (2) eliminating observations from the majority class – **undersampling**.

In this study, the dataset in question is unbalanced. Of the total observations (248956), approximately 86% are flights that did not arrive late (=0) and approximately 14% were reported as late (=1) as shown in Figure 8. It can be said that there is, approximately, a ratio of 6:1 where for each observation with delay there are six with no delay, making the existence of delay rare, i.e., a rare class or, more generally, an imbalanced class (Weiss, 2004). This is what happens most of the time in datasets with real data where the "normal" class is predominant and, only a small percentage refers to observations of the class of interest (Chawla, 2009; Chawla et al., 2002).

No.	Label	Count
1	0	213639
2	1	35317

Figure 8: Distribution of the dataset according to the dependent variable (ARR_DELAY) where 1 refers to an observation of a flight arrives late and 0 otherwise.
Source: Weka Software

An oversampling technique, SMOTE (**Synthetic Minority Over-sampling Technique**) was introduced by Chawla, Bowyer, Hall, & Kegelmeyer (2002) as a solution to this situation. It is one of the most used approaches due to its simplicity and effectiveness as well as for being able to improve the accuracy of the classifiers, and for that reason it was applied to this dataset through the ***SMOTE*** supervised instance filter provided in Weka.

It is used to create synthetic samples from the minority class instead of copies, similarly to a traditional random oversampling technique which can lead to overfitting problems where the model can adjust too much to training data by memorizing it and failing to correctly predict unknown data (Chawla, 2009; Hoens & Chawla, 2013). The algorithm, according to Hoens & Chawla (2013), "first selects a minority class *a* instance at random and finds its *k* nearest minority class neighbors. The synthetic instance is then created by choosing one of the *k* nearest neighbors *b* at random and connecting *a* and *b* to form a line segment in the feature space. The synthetic instances are generated as a convex combination of the two chosen instances *a* and *b*".

With the intent to see which is the best approach and check if there is a better approach besides the application of the oversampling technique – SMOTE –, it was also considered to implement a technique of undersampling, which could prevent common classes hiding rare classes (Weiss, 2004), through the ***SpreadSubsample*** supervised instance filter. This filter produces a random subsample of the dataset by specifying the maximum "spread" between the rarest and most common class, i.e., a ratio, assuming to be 1:1, reflects a dataset where for each observation with delay there is one with no delay, being a uniform distribution (University of Waikato, 2018). Observations from the majority class are ignored, and consequently, the training set becomes more balanced and the training process faster. In contrast, as a disadvantage, with this reduction of observations, useful information within this observations is neglected (Liu, Wu, & Zhou, 2009). For this reason, and in order to try to smooth the unbalanced problem in the dataset, a ratio of 2:1 is applied trying not to eliminate the presence of the majority class, that is, not to have a delay – most common case –, but also eliminating the massive difference of instances between the classes in the original dataset.

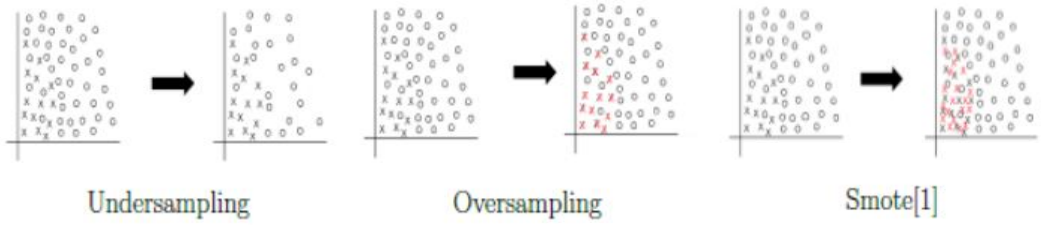


Figure 9: Sampling Techniques
Source: (Lachheta & Bawa, 2016)

This method generates two different approaches (one for the application of SMOTE, and another for the application of undersampling) to apply to the decisions taken in the course of this work.

3.1.4. Data Partition

Being a prediction problem, it is necessary to understand how a predictive model can succeed in unknown data, i.e., to perceive its generalization capacity. Its central purpose is to divide the data into unique subsets and then use a set(s) to estimate the model parameters - training data - and the remaining validation or test data - to validate the accuracy of the model. Always trying to manage the trade-off between a prediction error, and possible overfitting of the model with respect to the data, and the complexity of the model.

There are several approaches and methods for partitioning the data. The most straightforward, called **Holdout method** (Figure 10), is one that randomly divides the available data into two sets of training and testing or, three sets of training, validation, and testing, being adequate for large datasets.

In the first case, training data is used to train and model the algorithm, and the test data is used to access the performance of the training model. In the second case, the model is applied to the training set in order to learn from it, validated through the validation set where the predicted model error is estimated and where adjustments are made to model parameters and tested through the test set in order to evaluate the performance of the final classifier (Hastie, Tibshirani, & Friedman, 2009; Pennsylvania State University, 2017).

According to Kohavi (1995), the holdout method is a pessimistic estimator in the sense that only a portion of the data is presented to the algorithm for training which could represent a disadvantage when dealing with small datasets. It also presents a dilemma in choosing the size of the test data because if many observations are presented, the estimator is more skewed and, if few observations are presented, the confidence interval of the precision of the model will be higher. However, the author mentions that in this type of method it is usual to designate 2/3 of the data for training and only 1/3 for testing.

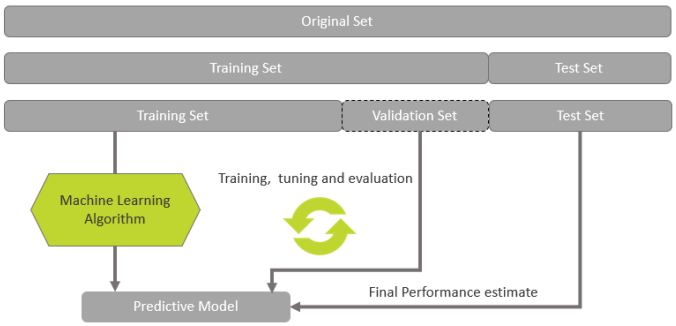


Figure 10: Holdout Method Training-Test and Training-Validation-Test
Source: Made by the author, adapted from (Wu & DataLab, 2016)

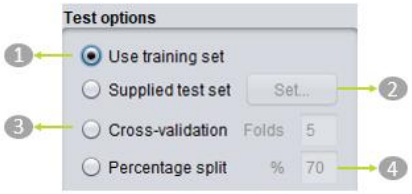


Figure 11: Data Partition options in Weka
Source: WEKA Software

In Weka, only four options are supplied as shown in Figure 11**Erro! A origem da referência não foi encontrada.**. In this study, there is only one dataset of examples labeled, namely the FlightData dataset. Through Weka and its four options, it is possible to (1) build a model on the FlightData dataset and apply it to the same dataset meaning that the testing would be done on the training set. It is also possible to (2) build a model upon the FlightData dataset and apply it to a second dataset if we had another separate file with examples not presented in training phase. It is also possible not having to decide which is the best to train and which is the best to test and to have the opportunity to do the two things, i.e., (3) divide the FlightData dataset into, for example, 5 equal folds, A, B, C, D, E. And then train in A, B, C, D and test on E, subsequently it will train on A, B, C, E and test on D and so on, testing each fold one time averaging the accuracy results of the five times that was done. And at last, (4) just divide the FlightData dataset by percentage attributing X to validation and the remaining of X to testing.

Applying the first option is not the best because the model will always be seeing the same data and will be overfitted, having a false precision. The second is not feasible because in this study there is only one dataset. The third option is recommended for small datasets, and in this study, we have a very large one. For that reason, the last options will be applied because of faster performance and for being the most suitable to the type of dataset in hand.

In this specific case, the original dataset will be divided into two datasets: Training and Test. For that reason, a training set is present to train and model the algorithm, and after the entire process, the test set is provided, not being used for training the model to estimate the accuracy of what was trained in unseen data.

Therefore, the hold-out method (equivalent to the option (4), percentage split) was chosen, and it follows the logic of Kohavi (1995) with a division of 70/30 for the test-training Holdout method. However, there is no rule on how much data is required for training and testing, always depending on the data noise as well as the complexity of the data to fit the model (Hastie et al., 2009).

3.2. DATA PRE-PROCESSING

Data pre-processing is one of the most important and at the same time difficult steps in the process. Pyle (1999), in his book *Data Preparation for Data Mining*, estimates that the task of data

preparation is equivalent to 60% of the time spent on a project. The need for a large amount of time and its importance in data quality due to noise, inconsistency or absence in very cases is what gives it great importance (De Ville, 2001; Witten et al., 2011a). Pyle (1999) further argues that the need to prepare and process the data helps prepare the researcher in the context of his/her knowledge of the data processed and, consequently, the design will be better and faster.

The main objective of this step is the GIGO concept (Garbage in, Garbage Out). Consists in minimizing the "garbage" that enters the model to minimize the amount of "garbage" that results from the model (Larose, 2005) showing that worked and meaningful data are a prerequisite in the production of effective models (Pyle, 1999).

Much of this work was done when constructing the database for creating the input dataset. However, this initial treatment was only made due to the need that arose in the integration of the data that was coming from different sources. At that stage, it was not solved any type of missing data and therefore, after correction of values and formats for integration in the database, some fields were still empty, and there was a need to pre-process the final dataset from the database. This is what is described along this sub-chapter because “the input to the data mining algorithms is assumed to be nicely distributed, containing no missing and incorrect values however, reality is much more “dirty” than the ideal and the data usually requires much preprocessing before any data mining algorithms can be applied” (S. Zhang, Yang, & Zhang, 2002).

3.2.1. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an approach to analyze the data and see its main characteristics as well as its behavior, either alone as together with the dependent variable, with no clear ideas of what to look for (Hand et al., 2001).

It has gained a position as the gold standard methodology to analyze data and was introduced by John W. Tukey (1961) where he defined data analysis as “procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data”. In agreement to Larose (2005), it allows the analyst to maximize insight of a data set, examine the attributes and their behavior, identify interesting subsets of the observations and develop an initial idea of possible associations between the attributes and the target variable.

For being a way to investigate variables through the look of histograms and the exploration of the relationships among sets of variables (Larose, 2005) more specifically, the independent variables by the dependent variable, a brief view of them was carried out, in the beginning, to see how the original data behaved.

It is important to note that each graph contains two types of information. One showing the frequency of each class by the dependent variable on the vertical axis. And the other, showing the relevance, in percentage, of each class of the dependent variable, by each class of the independent variable in analysis on the horizontal axis, beneficial to compare class importance. Some numerical variables were grouped into classes to ensure its simplicity and clarity in this current analysis.

The binary dependent variable, arrival delay, as already mentioned is not balanced in his original form, and for that reason, the flights with arrival delay only account for, approximately, 14% of the total flights in the database as can be seen in Figure 12.

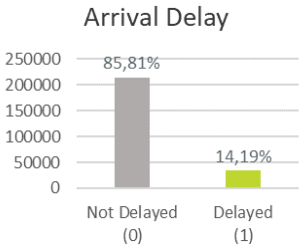
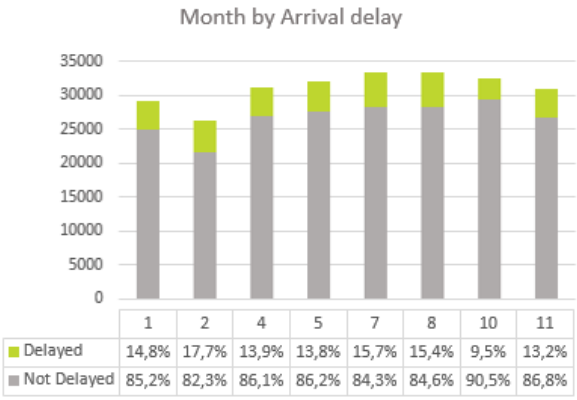


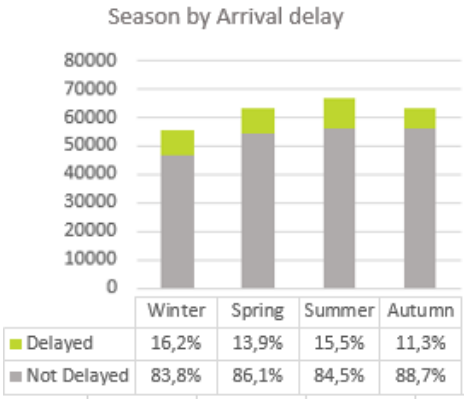
Figure 12: Dependent variable Arrival delay
Source: Made by the author

It can also be seen that the time of the year is also important. It can be seen through Figure 13 and Figure 14 that the month more relevant is February (17.7%) and July (15.7%) which have more weight on the arrival delay. For that reason, if one wants to see the months by seasons, it can be seen that the seasons that have more impact on the delay are Winter (16.2%) and Summer (15.5%) in accordance to the months more influent as also concluded by (Tu, Ball, & Jank, 2008b). This probably happens because of the weather in winter and due to the summer vacations when most people travel.



	1	2	4	5	7	8	10	11
Delayed	14,8%	17,7%	13,9%	13,8%	15,7%	15,4%	9,5%	13,2%
Not Delayed	85,2%	82,3%	86,1%	86,2%	84,3%	84,6%	90,5%	86,8%

Figure 13: Influence of Arrival delay on Month variable
Source: Made by the author



	Winter	Spring	Summer	Autumn
Delayed	16,2%	13,9%	15,5%	11,3%
Not Delayed	83,8%	86,1%	84,5%	88,7%

Figure 14: Influence of Arrival delay on Season (adaptation of Month variable)
Source: Made by the author

Looking at the days of the week it is curious that the days with more observations – flights – are Friday, equivalent to number 5, Thursday, number 4, and Monday, number 1 with 38 566, 38 343 and 36 583 respectively. However, being Friday the day with more flights it does not translate as the day which most contributes and influences the delay; instead it is Monday (16.7%) and Thursday (15.1%) as can be seen in Figure 15.

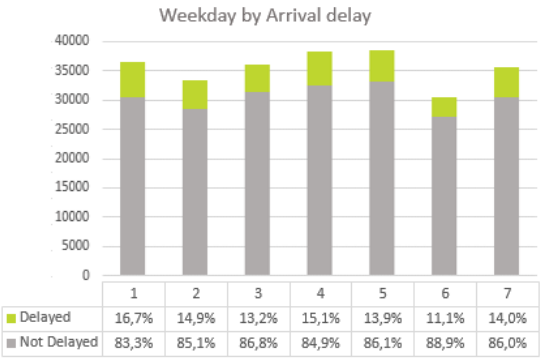


Figure 15: Influence of Arrival delay on Weekday variable
Source: Made by the author

It is interesting that, the concept of delay in arrival gain more strength as the expected duration of the flight increases. According to Figure 16, flights with expected duration in the range of 7h (420 minutes to 480, not included) have a higher percentage of delay on arrival comparing to flight with scheduled duration in the range of 8h (480 minutes to 540, not included) and 1h (60 minutes to 120, not included), 19.7% comparing to 15.8% and 14.6%, respectively.

Is important to note that most of the flights have a schedule duration time in the range of 1h and for that reason, the high percentage affecting delay arrival could be affected by other elements besides just the flight duration such as baggage, maintenance, cabin crew, along with others.

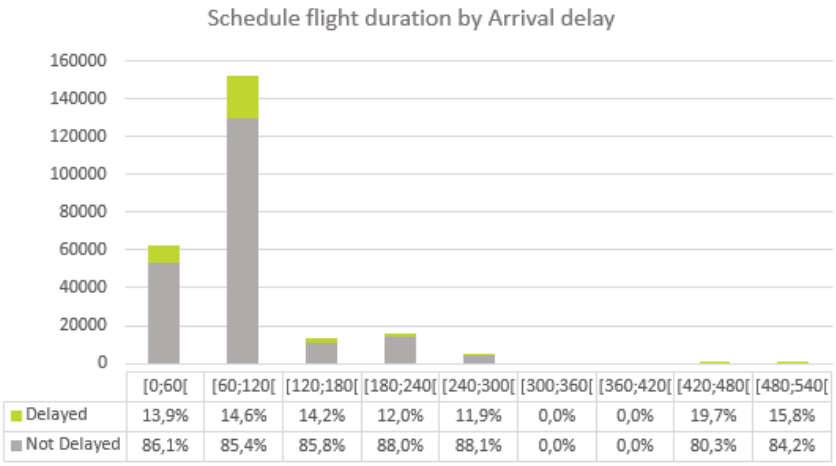
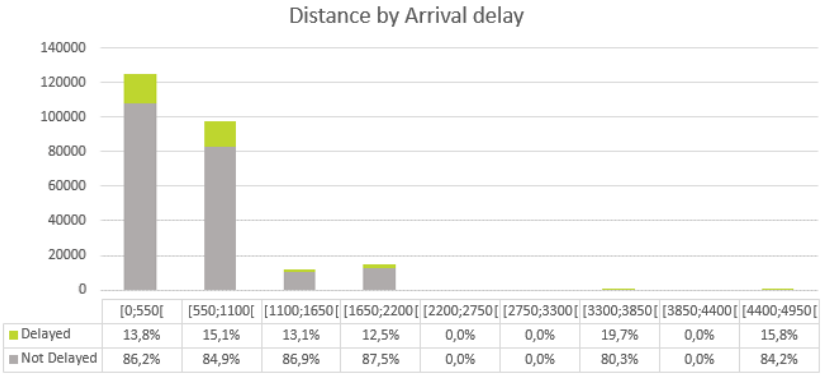


Figure 16: Influence of Arrival delay on Schedule flight duration variable
Source: Made by the author

The same can be seen in the distance between origin and destination of a flight. As illustrated in Figure 17 a higher distance reflects a higher percentage of arrival delays. Distances of 3300 miles to 3849 have 19.7% of delays compared to the 13.8% of delays on the interval of distance more frequent in the database of flights, from 0 to 550 miles.

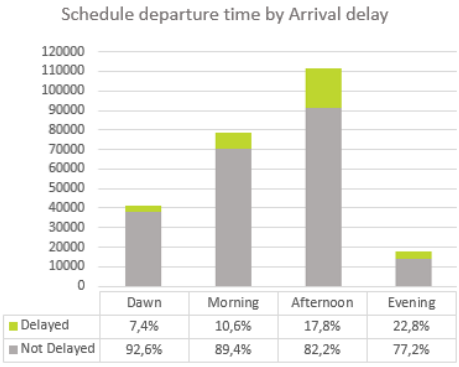


	[0;550[	[550;1100[	[1100;1650[	[1650;2200[	[2200;2750[	[2750;3300[	[3300;3850[	[3850;4400[	[4400;4950[
Delayed	13,8%	15,1%	13,1%	12,5%	0,0%	0,0%	19,7%	0,0%	15,8%
Not Delayed	86,2%	84,9%	86,9%	87,5%	0,0%	0,0%	80,3%	0,0%	84,2%

Figure 17: Influence of Arrival delay on Distance variable
Source: Made by the author

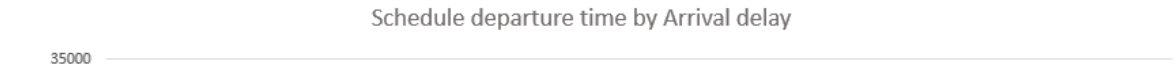
When looking at the schedule and real departure time when grouping the hour intervals in phases of the day, there is a part that has more delays in arrival in comparison to the others. There were represented four phases of the day for a better understanding of the behavior of the arrival delay by the course of the day. The first, dawn, referent to the intervals of minutes from the 0 minutes (00:00h) till 419 minutes (06:59h). The second one, morning, from 420 minutes (07:00h) till 719 minutes (11:59h). The third, afternoon, from 720 minutes (12:00h) to 1139 minutes (18:59h). And at least, evening, from 1140 minutes (19:00h) till 1439 minutes (23:59h).

Around 56% of delays at arrival occurs in flights that are designated to depart in the afternoon interval. Although that in the evening phase there is a higher percentage of delay comparing to the afternoon, 22.8% and 17.8% accordingly (Figure 18).



	Dawn	Morning	Afternoon	Evening
Delayed	7,4%	10,6%	17,8%	22,8%
Not Delayed	92,6%	89,4%	82,2%	77,2%

Figure 18: Influence of Arrival delay on Schedule departure time variable (in parts of the day)
Source: Made by the author



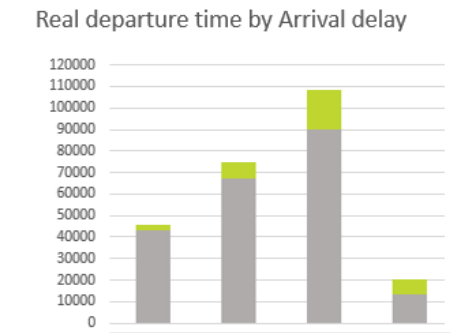
	[0;60[	[60; 120[	[120; 180[	[180; 240[	[240; 300[	[300; 360[	[360; 420[	[420; 480[	[480; 540[	[540; 600[	[600; 660[	[660; 720[	[720; 780[	[780; 840[	[840; 900[	[900; 960[	[960; 1020[	[1020; 1080[	[1080; 1140[	[1140; 1200[	[1200; 1260[	[1260; 1320[	[1320; 1380[	[1380; 1440[
Delayed	11,3%	15,5%	0,0%	0,0%	0,0%	4,8%	8,0%	9,4%	10,7%	10,8%	11,0%	11,7%	13,2%	14,8%	15,4%	17,0%	18,6%	21,8%	24,4%	23,0%	24,7%	32,6%	16,8%	17,5%
Not Delayed	88,7%	84,5%	0,0%	0,0%	0,0%	95,2%	92,0%	90,6%	89,3%	89,2%	89,0%	88,3%	86,8%	85,2%	84,6%	83,0%	81,4%	78,2%	75,6%	77,0%	75,3%	67,4%	83,2%	82,5%

Figure 19: Influence of Arrival delay on Schedule departure time variable

Source: Made by the author

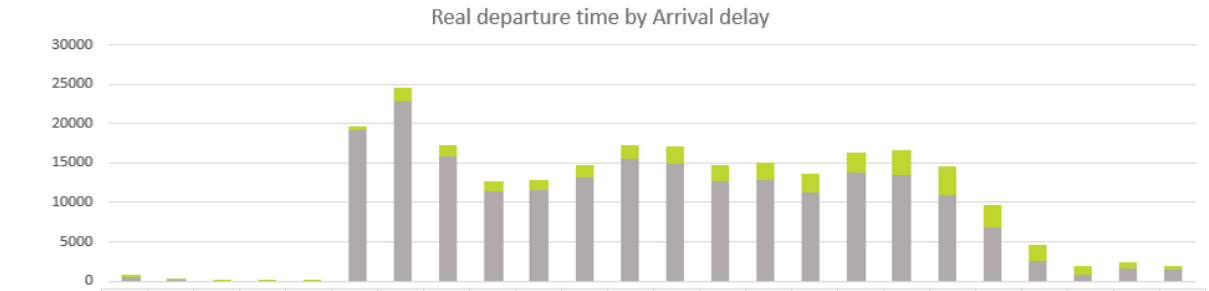
When seeing the real departure delay, it is noted that the proportion of flights that are delayed in the arrival increases starting from 18h (1080 minutes) till the rest of the day (Figure 21).

For that reason, there is a higher percentage of delayed flights in the evening (22.8%, in the scheduled departure time (Figure 18), comparing to 34.8%, in the real departure time (Figure 20)). This higher percentage of delayed flights in the evening interval is because the real departure time delays account for some flights that depart at the same hour of scheduled departure time but arrive late. And also by accounting the ones that departure at a different hour of the schedule one (more early or later) but also arrived late.



	Dawn	Morning	Afternoon	Evening
Delayed	5,6%	10,0%	16,8%	34,8%
Not Delayed	94,4%	90,0%	83,2%	65,2%

Figure 20: Influence of Arrival delay on Real departure time variable (in parts of the day)
Source: Made by the author



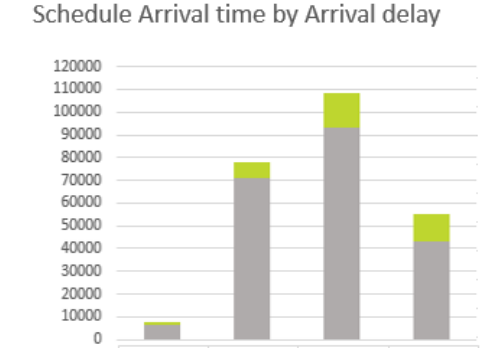
	[0;60[	[60; 120[	[120; 180[	[180; 240[	[240; 300[	[300; 360[	[360; 420[	[420; 480[	[480; 540[	[540; 600[	[600; 660[	[660; 720[	[720; 780[	[780; 840[	[840; 900[	[900; 960[	[960; 1020[	[1020; 1080[	[1080; 1140[	[1140; 1200[	[1200; 1260[	[1260; 1320[	[1320; 1380[	[1380; 1440[
Delayed	35,5%	34,6%	100,0%	100,0%	100,0%	2,1%	7,0%	8,6%	10,1%	9,9%	11,1%	10,5%	12,8%	13,8%	14,6%	17,5%	15,7%	19,3%	24,3%	29,3%	45,1%	56,1%	28,5%	24,5%
Not Delayed	64,5%	65,4%	0,0%	0,0%	0,0%	97,9%	93,0%	91,4%	89,9%	90,1%	88,9%	89,5%	87,2%	86,2%	85,4%	82,5%	84,3%	80,7%	75,7%	70,7%	54,9%	43,9%	71,5%	75,5%

Figure 21: Influence of Arrival delay on Real departure time variable
Source: Made by the author

It is reasonable to conclude that the flights more propitious to have a delay in arrival are (1) the ones that are scheduled to depart in the evening and in the afternoon and (2) the ones that the real hour of departure is in the evening interval. This is contrary to what is said by (Tu et al., 2008b) where the delays are higher in the middle of the day and lower in the evening.

In the case of the scheduled arrival time, it is noticed that around 42% of total delays at arrival occurs in flights scheduled to arrive in the afternoon interval, and 35% occur in flights scheduled to arrive in the evening interval. However, it is in the evening that a higher percentage of flights are delayed, 22.5% comparing to 13.8% in the afternoon (Figure 22).

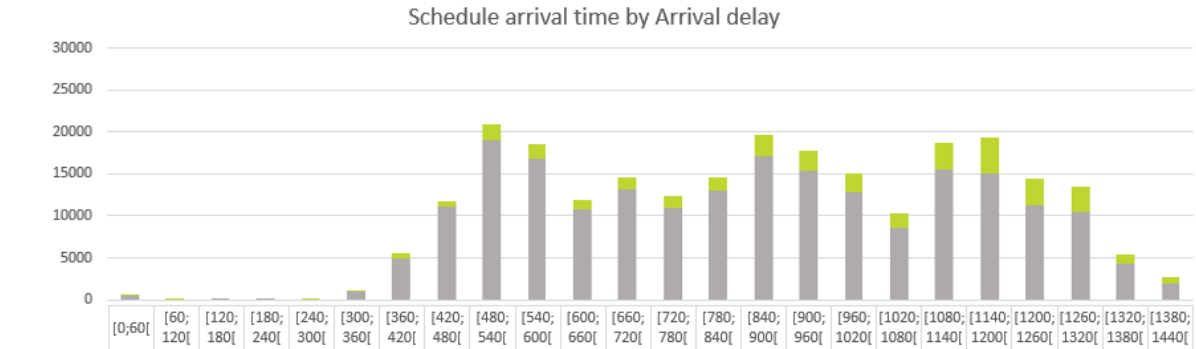
When analysing all intervals of hours individually (Figure 23), it is important to mention that in the interval



	Dawn	Morning	Afternoon	Evening
Delayed	10,9%	9,1%	13,8%	22,5%
Not Delayed	89,1%	90,9%	86,2%	77,5%

Figure 22: Influence of Arrival delay on Schedule arrival time variable (in parts of the day)
Source: Made by the author

between 60 minutes (01:00h) and 119 minutes (01:59h) is where the delay is most seen, i.e., in this interval 34.6% of the observations experience a delay in the arrival. This also happens in the previous interval from 0 minutes to 59 minutes and also in the intervals in the later periods of the day but with a slightly smaller percentage of delay in arrival.



	[0;60[	[60; 120[	[120; 180[	[180; 240[	[240; 300[	[300; 360[	[360; 420[	[420; 480[	[480; 540[	[540; 600[	[600; 660[	[660; 720[	[720; 780[	[780; 840[	[840; 900[	[900; 960[	[960; 1020[	[1020; 1080[	[1080; 1140[	[1140; 1200[	[1200; 1260[	[1260; 1320[	[1320; 1380[	[1380; 1440[
Delayed	21,5%	34,6%	0,0%	0,0%	27,3%	10,6%	9,2%	6,0%	9,4%	9,5%	10,2%	9,8%	12,0%	10,5%	12,4%	13,3%	14,4%	16,6%	17,4%	22,5%	21,8%	22,6%	22,4%	26,6%
Not Delayed	78,5%	65,4%	100,0%	100,0%	72,7%	89,4%	90,8%	94,0%	90,6%	90,5%	89,8%	90,2%	88,0%	89,5%	87,6%	86,7%	85,6%	83,4%	82,6%	77,5%	78,2%	77,4%	77,6%	73,4%

Figure 23: Influence of Arrival delay on Schedule arrival time variable
Source: Made by the author

Regarding the origin from which a flight departs it is possible to observe two things.

The first is that through Figure 24, where each cardinal location is shown by the percentage of airports that constituted them, the southeast and east have a higher percentage of cases because those regions have the highest amount of airports.

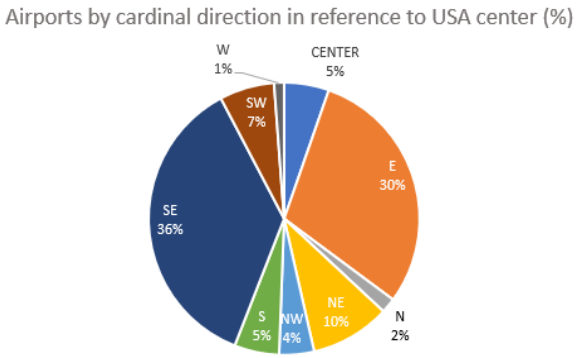
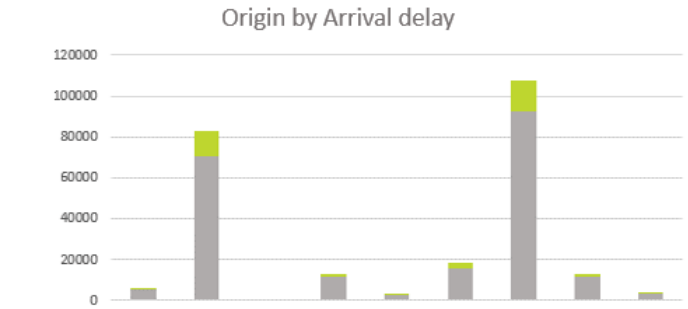


Figure 24: Distribution of airports accordingly to cardinal directions
Source: Made by the author

Secondly, it is interesting that when grouping the airport origins by cardinal location the airports located in the south are the ones with a higher percentage of delays when arriving in ATL (16%) (Figure 25) although 42% of delays occur in flights that originate in the southeast of USA.

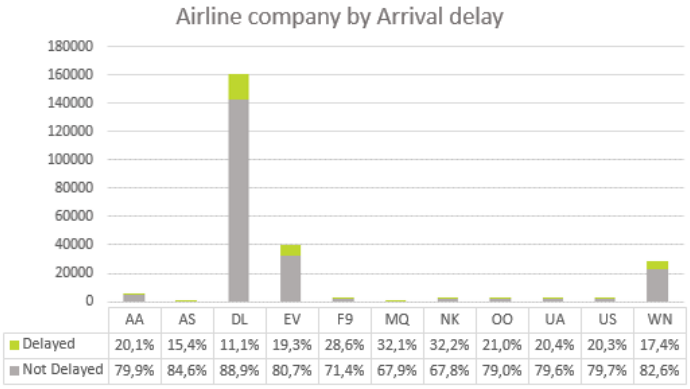


	CENTER	E	N	NE	NW	S	SE	SW	W
Delayed	15,1%	14,8%	9,7%	12,4%	11,8%	16,0%	13,9%	12,9%	10,7%
Not Delayed	84,9%	85,2%	90,3%	87,6%	88,2%	84,0%	86,1%	87,1%	89,3%

Figure 25: Influence of Arrival delay on Origin variable
Source: Made by the author

About the airline company operating the flight is known that 50% of the total delays are composed of flights operated by DL (Delta Air Lines Inc.). However, just 11.1% of flights operated by them are

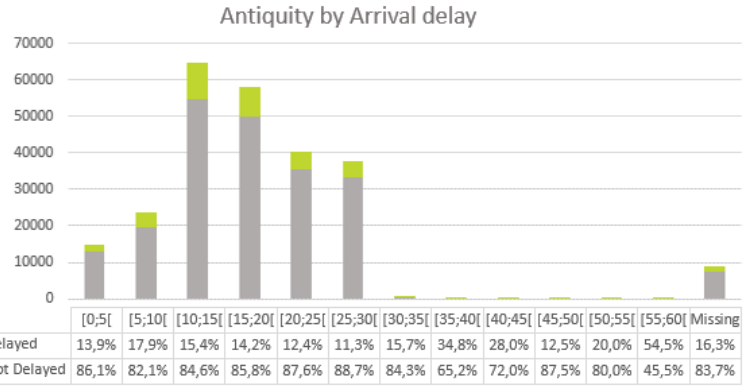
delayed in contrast to the 32.2% and 32.1% of flights delayed in the ones operated by NK (Spirit Airlines) and MQ (Envoy Air Inc.) respectively (Figure 26).



	AA	AS	DL	EV	F9	MQ	NK	OO	UA	US	WN
Delayed	20,1%	15,4%	11,1%	19,3%	28,6%	32,1%	32,2%	21,0%	20,4%	20,3%	17,4%
Not Delayed	79,9%	84,6%	88,9%	80,7%	71,4%	67,9%	67,8%	79,0%	79,6%	79,7%	82,6%

Figure 26: Influence of Arrival delay on Airline company variable
Source: Made by the author

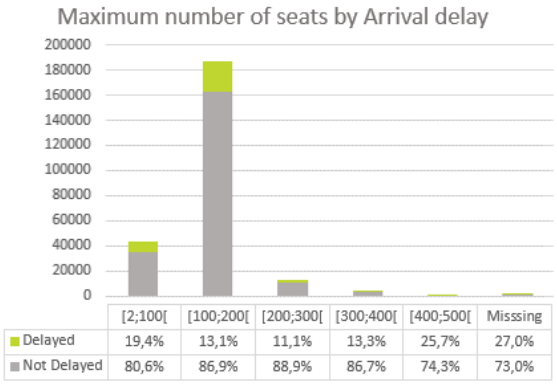
Regarding the airplane operating a specific flight, as concerns to the antiquity of an airplane there is a 28% of representation of the total delay in the ones with ten to fourteen years old. It is recognized that, with the assistance of Figure 27, in flights operating with airplanes within an antiquity between thirty-five to thirty-nine years old there are 34.8% delayed, the second interval with most delayed flights, being the first the ones within fifty-five and fifty-nine years old where 54.5% of flights are delayed in arrival.



	[0;5[	[5;10[	[10;15[	[15;20[	[20;25[	[25;30[	[30;35[	[35;40[	[40;45[	[45;50[	[50;55[	[55;60[	Missing
Delayed	13,9%	17,9%	15,4%	14,2%	12,4%	11,3%	15,7%	34,8%	28,0%	12,5%	20,0%	54,5%	16,3%
Not Delayed	86,1%	82,1%	84,6%	85,8%	87,6%	88,7%	84,3%	65,2%	72,0%	87,5%	80,0%	45,5%	83,7%

Figure 27: Influence of Arrival delay on Antiquity variable
Source: Made by the author

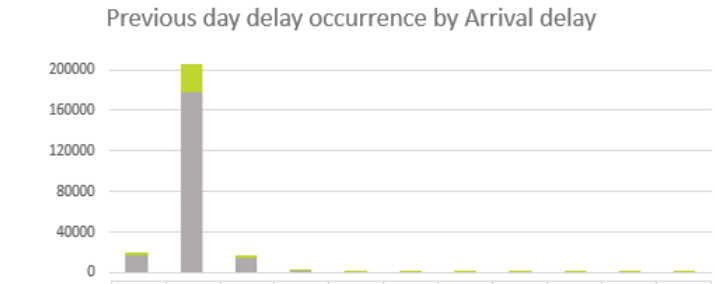
Related to the maximum of seats of the airplane operating a flight, it is noticed that the ones operated with a number of seats, in this case, between four hundred (400) and four hundred ninety-nine seats (499), have a higher percentage of delayed flights, 25.7% (Figure 28).



	[2;100[	[100;200[	[200;300[	[300;400[	[400;500[	Misssing
Delayed	19,4%	13,1%	11,1%	13,3%	25,7%	27,0%
Not Delayed	80,6%	86,9%	88,9%	86,7%	74,3%	73,0%

Figure 28: Influence of Arrival delay on Maximum number of seat variable
Source: Made by the author

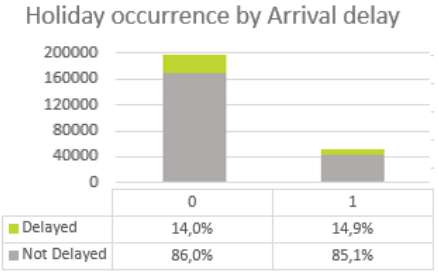
Regarding passengers' volume and existent traffic problematic, it is seen, with the assistance of Figure 29, that the greater the number of delays/ cancellations in the origin of the flight, on the day before, the greater the percentage of delayed flights.



	[0;1[	[1;100[	[100; 200[	[200; 300[	[300; 400[	[400; 500[	[500; 600[	[600; 700[	[700; 800[	[800; 900[	Missing
Delayed	12,6%	13,6%	19,5%	24,3%	29,5%	37,1%	22,1%	36,3%	32,1%	60,0%	10,6%
Not Delayed	87,4%	86,4%	80,5%	75,7%	70,5%	62,9%	77,9%	63,7%	67,9%	40,0%	89,4%

Figure 29: Influence of Arrival delay on Previous day delay occurrence variable
Source: Made by the author

In the interval from 800 minutes to 899 minutes, there are 60% of flights that experience delay on arrival comparing to the interval that represents 79% of the total delayed flights, from 1 minute to 99 minutes, where only 13.6% of flights are delayed. It is also seen in Figure 30 that around 14.9% of flights that presence a holiday in the same day of the flight or in a buffer of three days suffers a delay in arrival, against the 14% that do not presence a holiday but also have a delay in arrival.



	0	1
Delayed	14,0%	14,9%
Not Delayed	86,0%	85,1%

Figure 30: Influence of Arrival delay on Holiday occurrence variable
Source: Made by the author

In what weather variables are concerned, it is important to mention that there is an increase in the percentage of delayed flights in comparison to the ones where the temperature in the moment of observation is of extreme values, either positive and negative (see Annex 7 from Annexes chapter).

When observing the precipitation, it can be noticed that when the precipitation is in the range of 10 to 50 millimeters per hour, there is a higher percentage of delayed flights although few flights are occurring during this type of conditions. Nonetheless, the ones that occur in these terms have a higher percentage of delay comparing to the non-delayed, than the ones occurring in the other intervals (consult Annex 8 from Annexes chapter).

By analyzing the wind variables, it is detectable that there is a higher percentage of delays when the speed of the wind is higher (Annex 9 from Annexes chapter). The interval of 32 to 39 miles per hour is the one with the most records of delays in three phases of the variable (at the origin at the scheduled departure time; at the destination at the scheduled time of departure; and, at the destination at the scheduled arrival time). There is no record of higher values registered being an exception a record that registers a wind speed of more than 73 miles per hour (that is a specific case on the 14th of August in the airport origin of Myrtle Beach, South Carolina). In this specific case a wind speed of 166.9 mph the equivalent to 268.5 km/h was reported which leads to further research on the NAS website to understand what happened. No registry or mention of any major event was found, and, because of that, this type of observation will be considered an outlier.

The visibility variables show that the flights that contributes the most to the delay are the ones with visibility between 10 to 11 miles, being also the ones that constitute the non-delayed majority. As the visibility decreases, the percentage of delayed flights increases, in contrast to the ones that are not delayed as expected, nonetheless they are always inferior to the non-delayed ones (Annex 10 from Annexes chapter).

3.2.2. Missing Values

Missing values are known as data that are “missing for some (but not all) variables and some (but not all) cases” (Allison, 2001) in the database. They are critical to the management of the data because they can represent an issue in data quality and compromise the interpretation of data. It must be emphasized that replacing missing values is a gamble, and the benefits must be weighed against the possible weakness of the results (Larose, 2005).

Different reasons can cause them, such as equipment malfunctions or failures, inability to collect an observation, and so on (Batista & Monard, 2002). The lack of some observations in some variables can reduce the sample size, and as a result, the precision may be negatively affected, the statistical power weakened, and the parameter estimates could be biased (Soley-bori, 2013). For that reason, there is a need for a good understanding of this phenomenon and for appropriate measures to deal with it.

Soley-bori (2013) states that to deal with missing data requires a meticulous investigation of the data to be able to identify and understand the missing data to decide on how to treat data applying the technique that suits the most for why data is missing.

There are various ways to treat missing data. A common one is to ignore the observations or variables where missing data is present. This kind of approach can be risky because the patterns of missing data may be systematic.

Another way could be deleting them which could lead to bias, but also because deleting the observation or variables with missing data would omit the rest of the information for the remaining variables or the other observations, respectively, only because a few values are missing.

To escape from the drastic solution previously presented and when, it is possible to solve it by other means, measures less drastic are taken. This type of methods state that the missing values could be substituted according to some criteria such as (1) replacing the missing value by a constant, specified by the analyst; (2) replacing the missing value with the mean (in case of numerical variables) or the

mode (in the case of categorical variables) although it is argued that the mean may not always be the best choice for what constitutes the best value; this is because substituting missing values by the mean will turn the statistical inference overoptimistic since measures of spread will be reduced artificially; (3) replacing the missing values with a value generated randomly from the variable distribution, being a better method than the previous one (Larose, 2005); and, at last (4) replacing the missing values with imputed values based on other characteristics of the observation (Han & Kamber, 2011).

In this particular study, only some variables have missing values, and how they will be treated vary according to the reason why they have missing values. In the table below (Table 8) the variables which have missing values and the representative percentage concerning the total of instances are represented.

Variable	*# of Missing Values*	*% of Missing Values*
ANT	8873	4
MAN	1405	1
MOD	1405	1
MAX_SEATS	1464	1
PREV_DAY_DELAY_OCURR	1021	0
TEMP_ORIGIN	940	0
TEMP_DEST_SCHED_DEP_TIME	237	0
TEMP_DEST	204	0
PRECIP_ORIGIN	852	0
PRECIP_DEST_SCHED_DEP_TIME	237	0
PRECIP_DEST	200	0
WIND_ORIGIN	1335	1
WIND_DEST_SCHED_DEP_TIME	433	0
WIND_DEST	386	0
VISIB_ORIGIN	1005	0
VISIB_DEST_SCHED_DEP_TIME	237	0
VISIB_DEST	200	0
EVENT_ORIGIN	217477	87
EVENT_DEST_SCHED_DEP_TIME	210125	84
EVENT_DEST	210150	84

Table 8: Variables with missing values in the FlightData dataset
Source: Made by the author

There is a reason for every missing value and mainly is the unavailability of the data in the sources. The cases of missing values can be justified by the missing values in other variables or for being missing just for some reason not explained by another variable with observations missing. For example, variables of weather are missing because equipment do not report information for certain hours in certain airports. In other cases, the weather variables are not missing, but the variable of present weather condition (event) is missing, and this is because many times occurrences of this type of information are not reported on METAR in the sense that nothing worthwhile reporting happened.

That being said, in this research, the treatment carried out in the variables mentioned above was as follows:

- For the observations with no values about the airplane information such as antiquity, manufacturer, model, and maximum number of seat variables the approach that was taken was to remove them from the study through the ***RemoveWithValues*** filter. It does not make sense to replace them with values that do not match the real characteristics of the airplane and, as it represented a larger scope than the weather information (percentage of missing data), it will not have the same treatment. This is because it could impact the dataset and deviate it from the true values (a consequence of replacing observations with a mean or mode, depending on the variable type);
- For the observations where the variable of the presence of delay in the previous day was missing, only the case of January 1st (because of the lack of information of the previous day in the dataset downloaded since the previous day was not part of the year of 2015), the ***RemoveWithValues*** filter was applied;
- For all observations where no weather variable was available, because of a lack of records close to the source, imputation was applied through the ***ReplaceMissingValues*** filter. This decision was made so as not to lose observations with information that could be important for other variables and, since imputation did not involve many cases, the impact was considered to be low;
- In the case of the present weather information (event variable) there was a considerable amount of missing values. That probably happened because of equipment failure in capturing the information, and for that reason, the three variables concerning this topic were eliminated through the ***Remove*** filter.

3.2.3. Outliers

Hawkins (1980) defines an outlier as "an observation which deviates so much from other observations as to arouse suspicious that it was generated by a different mechanism". For Grubbs (1974) "an outlying observation, or "outlier" is one that appears to deviate markedly from other members of the sample in which occurs". Is also defined as a single, or very low frequency, occurrence of the value of a variable that is far away from the extent of the values of the variable (Pyle, 1999). In other words, outliers are extreme values that lie near the limits of the data range or go against the trend of the remaining data (Larose, 2005). They can be represented as individual occurrences and, sometimes, in clusters of consecutive values of the same order of magnitude but also as a group situated beyond the range of the other values (Pyle, 1999) as represented in Figure 31.

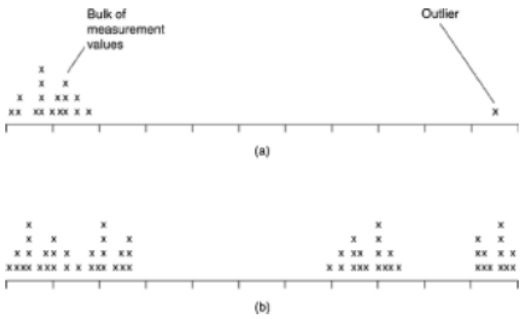
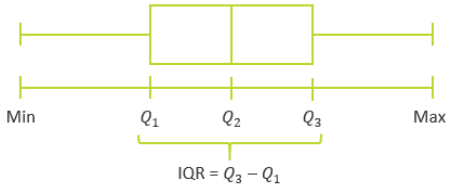


Figure 31: Examples of Outliers: as an individual value (a) and as clusters of values (b)
Source: Withdrawn from (Pyle, 1999)

When presented in the input data they could skew and mislead the training process of machine learning algorithms resulting in large training times, less accurate models and poor and unstable results because certain methods are sensitive to the presence of outliers (Larose, 2005).

For those reasons, it was necessary to identify which observations were outliers. Outliers could be caused by devices malfunction, fraudulent behavior, human error, natural deviations, among others. Nevertheless, one should be careful when dropping outliers. It is essential to investigate the nature of the outlier before deciding on what to do because each case is different and there is no right way to do so.

With the help of Weka and its filter of ***InterquartileRange,*** it was possible to identify what observations were **outliers** and **extreme values** based on interquartile ranges. This filter adds two new attributes whether the values of instances can be considered outliers or extreme values. As can be seen in Figure 32 the filter considers an observation to be an extreme value if they exceed the upper quartile (Q3) or fall below the lower quartile (Q1) by the product between the extreme value factor (user-specified) and the interquartile range (IQR). And, to be an outlier, the ones that are not extreme values but exceed the upper quartile (Q3) or fall below the lower quartile (Q1) by the product of the outlier factor (user-specified) and the interquartile range (IQR) (Witten et al., 2011b). Correspondingly to the default settings of Weka, and the ones taking in consideration when applying the filter, the value of the outlier factor is three (OF=3), and for extreme values, the factor is two times the outlier factor (EVF=OF*2).



Extreme value when,

$$x > Q_3 + EVF * IQR$$

or

$$x < Q_1 - EVF * IQR$$

Outlier when,

$$Q_3 + OF * IQR < x \leq Q_3 + EVF * IQR$$

or

$$Q_1 - EVF * IQR \leq x < Q_1 + OF * IQR$$

Figure 32: Definition of Extreme Value and Outlier using *InterquartileRange* filter
Source: Made by the author, adapted from Weka *InterquartileRange* ObjectEditor Information
Where Q_1 = 25% quartile; Q_2 = 50% quartile (median); Q_3 = 75% quartile; IQR = Interquartile Range; OF = Outlier Factor; EVF = Extreme Value Factor

The application of the filter resulted in two new attributes: one with extreme value information and another with outlier information, both of binary type. The next step consisted of eliminating the records where the observations were considered to be extreme values and then repeat it to the observations considered to be outliers, through the filter ***RemoveWithValues***.

For a comparison of performance between the decision of eliminating the outliers and extreme values, and the decision not to eliminate them, it was considered, for each created dataset, regarding the use and non-use of the variable *Dep_Delay* and *Real_Dep_Time*, and the application of SMOTE

and Undersampling, the creation of two other datasets: one contemplating the removal of outliers and another not considering them.

3.3. DATA TRANSFORMATION

Besides all the steps mentioned above, other steps could be taken to improve the performance and success of the model. Witten, Frank, and Hall (2011b) state that these other steps could be considered "a kind of data engineering—engineering the input data into a form suitable for the learning scheme chosen and engineering the output to make it more effective". It is also mentioned in their book that, in the state of the art, there is no guarantee that this kind of steps will work but it is reinforced that in this area, a trial and error approach is probably the best.

For our purposes, and having in mind the aiming of the study, there are two ways, among others, to adapt the input and make them more susceptible to the learning methods that will be developed inside this sub-chapter: normalization and attribute selection.

3.3.1. Normalization

There is a need of data miners to normalize numerical variables with the intent of standardizing the scale of the effect that each variable has on the results. This translates into an increase of the value of the model by equaling the weights of different scales and meanings of variables. There are several techniques to do this step, and the most popular methods are z-score and min-max, being the last the one used in this study, on all interval variables through the ***Normalize*** filter in Weka.

The normalization of data through the **min-max method** is a procedure that rescales variables in a range between 0 and 1 where the largest value of each variable is 1, and the smallest value is 0. The normalized field value (X*) is obtained by subtracting the minimum value from the original field value (X) and dividing it by the difference between the maximum and minimum values (Larose, 2005).

$$X^* = \frac{X - min(X)}{max(X) - \min(X)}$$

Equation 1: Min-Max Normalization Technique
Source: Made by the author, adapted from (Larose, 2005)

It is considered to be a suitable technique when the data has varying scales, which is the case iin this study. Its importance is related to the fact that some algorithms are more sensitive to these discrepancies in data ranges than others, which can result in a variable with higher values to have an inadequate influence on the results (Larose, 2005).

3.3.2. Variable Selection

A models' complexity can be a detrimental factor to a clear and easy understanding of it. This complexity results from the vast amounts of variables for algorithms to handle which most of the times are irrelevant or redundant to the problem having no importance in the dependent variable arrival delay (Han & Kamber, 2011). For such reason, these type of attributes were deleted in behalf of a better understanding, a faster execution, and a higher possibility of a better performance of the model, despite having a substantial computational cost (Saeys, Inza, & Larranaga, 2007).

For Witten, Frank, & Hall (2011b), it is clear that algorithms can already seek for the best attributes and ignore the irrelevant and redundant ones; nonetheless, their performance can often be enhanced by preselection.

Attribute selection, also known as variable selection, feature selection or variable subset selection "is the process of identifying and removing as much of the irrelevant and redundant information as possible" (M. A. Hall & Holmes, 2002). It can follow a set of steps in its implementation such as (1) the generation procedure, where subsets of features are generated for evaluation based on a given search method; (2) the evaluation function, to evaluate the subset that is being examined according to a certain attribute evaluator; (3) the stopping criteria, to help to decide when to stop the search, and; (4) the validation procedure, where a subset is checked for being valid or not (Dash & Liu, 1997).

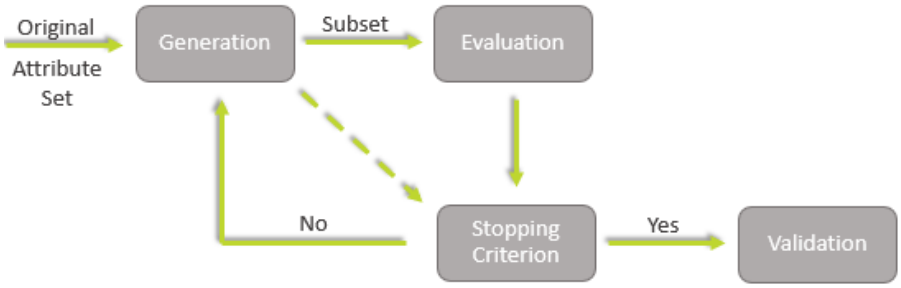


Figure 33: Attribute Selection Steps
Source: Made by the author, retrieved from (Dash & Liu, 1997)

The process of attribute selection is separated into two parts, the **attribute evaluator** that performs the evaluation function, and the **search method** that does the generation procedure.

In the first part, attribute evaluator, there are two types of evaluators as explained in *Data Mining: Practical Machine Learning Tools and Techniques* (2011b): the **attribute subset evaluator** and the **single-attribute evaluator** (Figure 34).

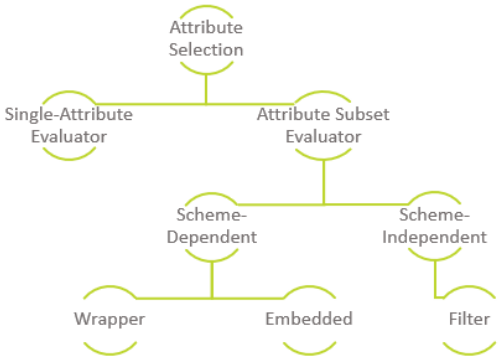


Figure 34: Attribute Selection Types
Source: Made by the author

The former takes a subset of attributes and returns a numerical measure that pilots the search. It has, as an asset, the fact that eliminates redundant and irrelevant attributes, and consequently, its search slow. This type of evaluators could be **scheme-independent** when they do not involve a classifier using only the search method and evaluating through a heuristic, i.e., selecting variables for a subset regardless the algorithm, based only in general measures as correlation, for example, and suppressing the variables with least interest for modeling. Nonetheless, they tend to select redundant variables because they do not account for relationships between variables; they are also known as **Filter methods** (Hamon, 2013)(Figure 35).

Contrarly to shceme-independent type exists the **Scheme-dependent** attribute subset evaluator type. It can be of **Wrapper method** type (Figure 36) (Kumar, Kongara, & Ramachandra, 2013) when the feature subset selection algorithm guides a search for a good subset using the induction algorithm itself as part of the function that is evaluating feature subsets (Kohavi & John, 1997). Or

Embedded method type (Figure 37) (Kumar et al., 2013), when trying to combine both advantages of the previous types presented, performing the variable selection by incorporating it with the algorithm directly, selecting the attributes that best contribute to the accuracy of the model when it is being trained (Witten et al., 2011b), and carrying out feature selection and classification at the same time (Hamon, 2013).

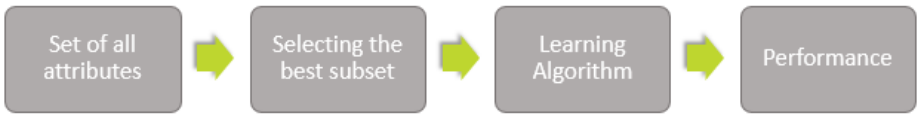


Figure 35:Attribute Subset Evaluator, Filter Method Selection
Source: Made by the author, adapted from (Hamon, 2013)

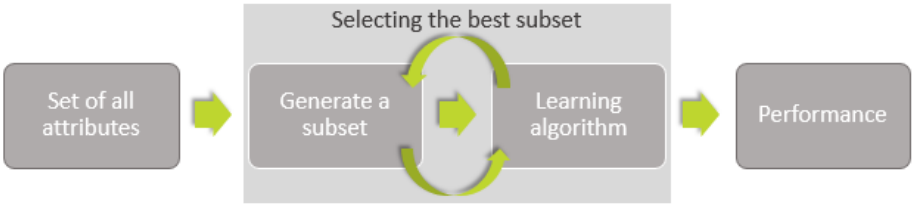


Figure 36: Attribute Subset Evaluator, Wrapper Method Selection
Source: Made by the author, adapted from (Hamon, 2013)

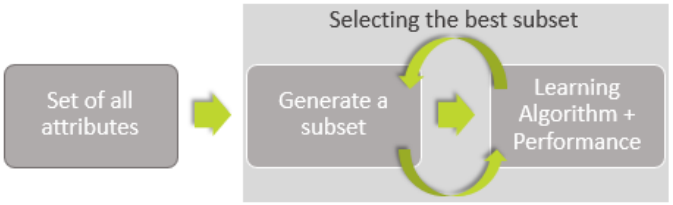


Figure 37: Attribute Subset Evaluator, Embedded Method Selection
Source: Made by the author, adapted from (Hamon, 2013)

The following attribute evaluator type, the single-attribute evaluator (Figure 38), represents an alternative to the slow performance of the former, and outputs an ordered list, composed by the number of attributes to keep and arranged by the importance of the attributes, in respect to a metric, through ranking. In its favor has its faster performance overweighed with the power to eliminate irrelevant attributes instead of irrelevant and redundant.

Figure 38: Single-Attribute Selection
Source: Made by the author

The second part, the search method, goes through the attribute space to find a good subset for the evaluator to measure its quality. It can be performed through the **BestFirst** and **GreedyStepwise** search methods, among others, for the attribute subset evaluators, or a **Ranker method** for the single-attribute evaluator (Witten et al., 2011b).

To see the performance of each one of the types of evaluators, it was selected one from each of them with the exception of wrapper and embedded methods. This is because acquiring a better

predictive accuracy has high needs of computational efforts and can select a subset of features that is biased to the predefined classifier (Tang, Alelyani, & Liu, 2014).

For the single-attribute evaluator, among others, the ***GainRatioAttributeEval,*** with the ***Ranker*** search method, maintaining the best variables according to the number that better improves the results, was chosen due to its ability to evaluate attributes by measuring their gain ratio with respect to the class and then ordering them according to their evaluation.

Concerning the attribute subset evaluator category the ***CfsSubsetEval*** (Correlation-based Feature Selection) was used, with the ***BestFirst*** search method being a greedy hill-climbing search augmented with a backtracking aptitude. It assesses the predictive capability of each attribute individually, along with the degree of redundancy between them promoting sets of variables highly correlated with the class variable but with the lowest intercorrelation (Witten et al., 2011b). It is a good evaluator because it takes into consideration the variables correlation that when used could cause instability and inaccurate results for the model (Larose, 2005).

3.4. DATA MINING

Being Data Mining, earlier mentioned in chapter 2 (Literature Review), the step of KDD where algorithms are used to extract patterns from the data translating it into information (Fayyad et al., 1996), in this subchapter, the main objective is to define and present the learning algorithms used.

DM has two main tasks: **Predictive modeling** (supervised learning) and **Descriptive modeling** (unsupervised learning). The former aims to learn decision criteria, making it possible to classify a new and unknown value of interest given known observation of other variables. The second aims to find hidden patterns in the data, summarizing data in ways that will guide to a higher understanding of the data (Hand et al., 2001; X. W. X. Wang, 2009).

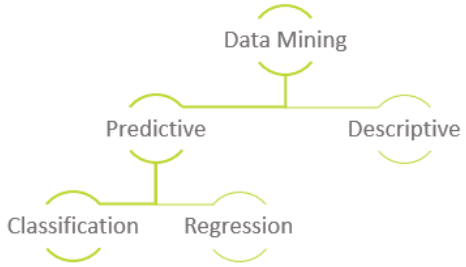


Figure 39: Data Mining Tasks
Source: Made by the author

As the interest of this work is to predict if flights are likely to be delayed or not, it is clear that the task inherent to this study is prediction. Furthermore, the predictive modeling task can be distinguished in two types of problems. The first, **regression problem**, if the value of interest to predict is continuous. And, the second, **classification problem**, if the field to predict is categorical, being what happens in this study due to the fact that the variable of interest "Arr_Delay" is a binary one, hence considered a binary classification.

With the interest of helping the predictive modeling task, there are several tools also known as learning algorithms. This learning algorithms, in the context of this study, were chosen regarding the three main articles in which this work relies on (Belcastro et al., 2016; Choi et al., 2016; Y. J. Kim et al., 2016), and also by providing a simple algorithm that could be easy to understand in contrast with the others. For that reason, the algorithms selected to applied were the ones with the best performance within each article, **Random Forests** (Belcastro et al., 2016; Choi et al., 2016) and **Multi-Layer Neural Networks** (Y. J. Kim et al., 2016). However, to also understand if, with more simplicity,

understandability, and ability to handle all types of variables, higher results could be achieved, the implementation of Decision Trees was also carried out.

Each learning algorithm has their parameters that, for a better result and understandability, needs to be wisely parametrized. This task is time-consuming, and it requires an excellent knowledge of each algorithm and the domain (Koblar, 2012). For this reason, and for the number of efforts needed to perform this task successfully, the default settings presented by Weka for each algorithm were used as a reference for this work.

3.4.1. Decision trees

Decision Trees (DT) are known as a "collection of decision nodes, connected by branches, extending downward from the root node until terminating in leaf nodes" (Larose, 2005) that could also be represented as sets of if-then rules improving the understandability. It is a method widely used for inductive inference (Mitchell, 1997) with a divide and conquer philosophy and advantages like the easy interpretation, the ability for processing a sort of data types, among others.

It works by classifying instances by sorting them down from the root to a leaf node that provides the classification of an observation. This is done by first selecting an attribute - based on a statistical test to determine how well it classifies the training examples- to place at the root node making one branch for each possible value. The entire process is then repeated recursively for each branch, using only the instances that reach the branch, where it is selected the best attribute again to test at that point of the tree. If at any time all instances of a node have the same classification, the development of that part of the tree must be stopped (Larose, 2005; Mitchell, 1997; Witten et al., 2011b). In a general definition Mitchell (1997) refers that decision trees "represent a disjunction of conjunctions of constraints on the attribute values of instances" where each path from the root to the leaf is a conjunction of attribute tests and the entire tree a disjunction of these conjunctions. Decision trees seek to obtain leaf nodes that are as pure as possible – each leaf only represents records within the same class – through the discrimination between classes (Larose, 2005).

Various algorithms have been developed for learning decision trees being variations of the core algorithm that applies a top-down, greedy search through the space of possible decision trees such as the ID3 algorithm – a basic algorithm for decision tree learning -, and its successor C4.5, an extension from the former (Mitchell, 1997). C4.5 came to handle shortcomings in ID3 algorithm. This is done by accepting both continuous and discrete variables; by using "pruning" to solve overfitting and improving the predictive accuracy removing sections of the tree that have no power to classify observations; by using an alternative measure for selecting attributes, the gain ratio, that takes into account the split information (term sensitive to how broadly and uniformly an attribute can split the data that prevent to select attributes with many uniformly distributed observations) and contrarily to the information gain, does not favor attributes with many values, over those with few if they are not a useful predictor over unseen instances; by handling missing data; and, by treating attribute with different weights when, in certain learning tasks, it is necessary (Mitchell, 1997; Quinlan, 1993).

In Weka, it is possible to apply this algorithm with the use of the tree classifier ***J48***, the equivalent to the C4.5 decision tree algorithm, adopting the default settings.

3.4.2. Random Forests

Random Forests is one of the most well-known algorithms of **Ensemble methods** that can be used for predictive modelling, either in classification or regression problems. Ensemble methods are a combination of two or more methods whose predictions are combined in an appropriate way, and that classify new data points by taking a vote of their predictions being often more accurate than the individual algorithms that constitutes them (Dean, 2014; Kumar et al., 2013; C. W. Wang, 2006).

They are based on the **Bagging (Bootstrap Aggregating) method** that has a great success and often outperforms single classifiers that make them (C. W. Wang, 2006). Bagging is based on the idea of generating multiple bootstrap training sets, with the same size and with replacement from the original one using each one to generate a classifier for incorporation in the ensemble. Classification of each unknown instance is done by majority vote (unweighted), where each individual classifier has equal weight and the winner is who has the majority of votes, or weighted where each base classifier has different voting power (Dean, 2014; Kumar et al., 2013; C. W. Wang, 2006; Witten et al., 2011b).

Random Forests are then considered as an ensemble of classifiers (Kumar et al., 2013) where, through the use of bagging, produce diversity among the predictors and also by growing innumerable classification trees and then recurring to voting and, as a final decision the majority prevails (Kumar et al., 2013). It assures the use of all variables - by splitting each node using the best, among a subset of variables randomly selected, at that node - and not only the most relevant – as it happens in decision trees where each node is split using the best split within all variables. For that reason, normally it encompasses a huge number of decision trees (Dean, 2014; Liaw & Wiener, 2002). It works by selecting a bootstrap sample for training; in the root node, a random sample of dependent variables is selected, and the best split is made by using that limited range of variables; in subsequent nodes, another random sample is performed, and the best split is applied. The tree grows this way until it reaches the largest possible size without being pruned. This process is restarted several times, always starting with a new bootstrap sample, being the final prediction a plurality vote or an average of predictions of all trees in the ensemble (Moisen, 2008).

The application of this algorithm in the Weka software is allowed by using the tree classifier ***RandomForest*** and by adopting the default settings.

3.4.3. Multilayer Neural Networks

A Multilayer Perceptron (MLP) is a type of artificial neural network that is, according to (Shalev-Shwartz & Ben-David, 2014), an algorithm inspired in the structure of neural networks existent in the human brain. It is based on the basic neural networks architectures that are composed by nodes (neurons) where each node receives an input signal – external information (environment) or received by other nodes - and an additional input included as the bias, a threshold element, connected via adjustable interconnection weights, processes through an activation or transfer function and converts the input to output – to pass to other nodes or to external outputs (environment)(Palit & Popovic, 2005; G. Zhang, Patuwo, & Hu, 1997). It is considered as a powerful tool for data classification, among other tasks. It can learn and generalize from observation data but, because it has a simple structure, a single perceptron alone cannot learn the sufficient amount of information to be efficient on solving more complex issues, being only capable of solving linearly separable

problems, which is a disadvantage when dealing with complexity (Mitchell, 1997; Palit & Popovic, 2005; Pinkus, 1999).

The MLP worked as a solution for this disadvantage being able to solve nonlinear problems. It is stated to be the most frequently used model in the class of artificial neural networks models despite of the variety of proposed neural networks structures that arose (Hand et al., 2001; Palit & Popovic, 2005; Pinkus, 1999). It differs from the traditional artificial neural networks by adding a hidden layer apart from the input and output layers that interconnect them (Palit & Popovic, 2005). Hand et al. (2001) explain that "the basic idea is that a vector of p input values is multiplied by a p × d1 weight matrix, and the resulting d1 values are each individually transformed by a nonlinear function to produce d1 "hidden node" outputs. The resulting d1 values are then multiplied by a d1 × d2 weight matrix (another "layer" of weights), and the d2 values are each put through a non-linear function. The resulting d2 values can either be used as the outputs of the model or be put through another layer of weight multiplications and non-linear transformations and so on". Palit & Popovic (2005) remarks that to construct a model with the computational capabilities of the MLP for solving the majority of complex practical problems there is only needed one hidden layer, and in rare cases, a few additional hidden layers will be needed.

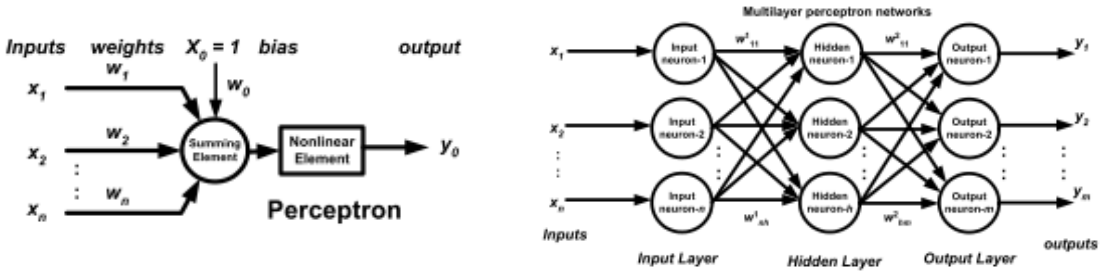


Figure 40: Artificial Neural Network and Multilayer Perceptron Architectures, respectively
Source: Withdrawn from (Palit & Popovic, 2005)

With the arrival of MLP also came the **backpropagation training algorithm**, a frequently used algortihm in real applications, that had a wide acceptance and the multilayer perceptron networks trained to learn using the backpropagation algorithm starts to be known as backpropagation networks (Mitchell, 1997; Palit & Popovic, 2005). It is essentially, a gradient descent method that in each iteration does a forward activation to produce an output and a backward propagation of the error computed (difference between the output and the target output) in order to adjust the weights from the output layer, through the hidden layers to the input layer, being done repeatedly until the solution meets the target value within a certain threshold (Basheer & Hajmeer, 2000; Lippmann, 1987), minimizing the squared error between network output and target value for the outputs (Mitchell, 1997).

The implementation of this type of algorithm in Weka is possible through the use of ***MultilayerPerceptron*** function classifier, that is accordingly to the information provided in Weka software, a classifier that uses backpropagation to classify instances using the sigmoid function as the nonlinear function.

As in other situations, some fields are user-specified such as the number of hidden layers, the learning rate, and the momentum. For the number of hidden layers, it will be used the default value where the number is defined by the sum of the number of attributes with the number of classes

divided by two. For the momentum, it is mentioned that a high value will reduce the risk that the network stays on local minimum and reduce the likelihood of instability in the search but it has the risk of going through the solution without recognizing it. Otherwise, small values will lead to a slow training. Values suggested range between [0.1;1.0] being the default value of Weka, 0.2, a good value (Basheer & Hajmeer, 2000). In respect to the learning rate, which determines the magnitude of weight changes (G. Zhang et al., 1997), if a high value is defined it will stimulate training by changing the weights significantly over iterations which could lead to a risk of overcoming an optimal solution. Contrarily, a small value guides the search in the direction of a global minimum, although slowly being enforced that it should stay the same throughout the training process. Values proposed vary from [0.1;10] being the Weka default value of 0.3 a reasonable estimate. It is also suggested that the sum of this two values (momentum and learning rate) is approximately 1 (Basheer & Hajmeer, 2000).

3.5. EVALUATION

After applying a set of algorithms, it is necessary to understand which ones perform best with the given data. The term "best" could be a quite complex and there is not a single definition of "best" when selecting an algorithm (Dean, 2014). In this task various measures can together lead to the choice of the "best" algorithm.

A variety of measurements is available such as the Confusion Matrix, Lift charts, ROC Curves, Recall and Precision, Cost Curves (Witten et al., 2011b), Gain, Akaike's Information Criterion, Bayesian Information Criterion and Kolmogorov-Smirnov (Dean, 2014).

In this work, relying on the results of Weka, and according to the problem under study, a classification task for a two-class problem will be used with the following measures:

- CCI – **Corrected Classified Instances**, also known as Accuracy (ACC), is the percentage of test instances correctly classified. Can have as a disadvantage that is not sensitive to the class distribution (problem of unbalanced datasets, necessary to take in account when interpreting results), representing a poor measure alone because it may result in high accuracy rates although the classifier may not be good;

$$CCI = ACC = \frac{TP + TN}{TN + FP + FN + TN} = \frac{TP + TN}{n}$$

Equation 2: Corrected Classified Instances Formula
Source: Retrieved from (Kononenko & Kukar, 2007a)

- PR - **Precision** is the proportion of instances that are legitimately true instances of a class divided by the total number of cases classified as being that class, i.e., estimates the portion of correctly classified examples that were classified as positive (Kononenko & Kukar, 2007a);

$$Precision = \frac{TP}{TP + FP}$$

Equation 3: Precision Formula
Source: Retrieved from (Kononenko & Kukar, 2007a)

- RE – **Recall** is the proportion of instances classified as a given class divided by the actual total of true values in that class. It can be defined as a relative frequency of correctly classified positives examples (Kononenko & Kukar, 2007a)

$$Recall = \frac{TP}{TP + FN}$$

Equation 4: Recall Formula
Source: Retrieved from (Kononenko & Kukar, 2007a)

- F - **F-Measure** is a measure that combines precision and recall.

$$F = \frac{2 * recall * precision}{recall + precision} = \frac{2 * TP}{2 * TP + FP + FN}$$

Equation 5: F-Measure Formula
Source: Retrieved from (Witten et al., 2011b)

- ROC Curve – corresponding to the **Area Under the ROC** (Receiver Operating Characteristics) curve (AUC) where the larger the area, the better the model and could be interpreted as the probability that the classifier ranks a randomly chosen positive instance above a randomly chosen negative one (Witten et al., 2011b). It is widely used in some research areas (Hand et al., 2001). Through the Figure 41 it can be seen that if the aim is to have the ROC curve as close as possible to the coordinates (0,1) then, the bigger the area under the ROC curve, the better the classifier. ROC curve illustrates the relationship between sensitivity (also known as recall) – a relative frequency of correctly classified positive examples -, and specificity – relative frequency of correctly classified negative examples (Kononenko & Kukar, 2007a)

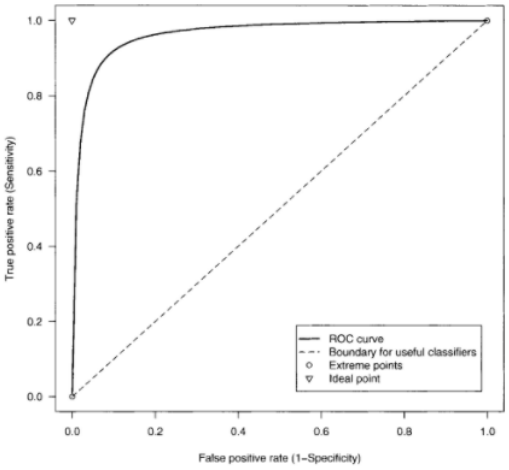


Figure 41: ROC Curve and Ideal Point
Source: Retrieved from (Kononenko & Kukar, 2007a)

- T - **Time** (in seconds) as a measure of time-consuming performance of the algorithm in building and testing the model as a metric for helping as a tiebreaker when needed.

Elements in formulas above are expressed as,

$TP - number\ of\ true\ positive\ observations;$
$FP - number\ of\ false\ positive\ observations;$
$TN - number\ of\ true\ negative\ observations;$
$FN - number\ of\ false\ negative\ observations;$
$n - total\ number\ of\ observations.$

Optimal results can be regarded as values close to 1 for all measurements except for the Corrected Classified Instances where the objective is values close to 100, and for the time measure where the smaller, the better.

4. RESULTS AND DISCUSSION

The only source of knowledge is experience.

– Albert Einstein (1879-1955)

In this work, three classifiers over different approaches using Weka software v. 3.8.1 were evaluated on a computer with an Intel® Core™ i5-6200U processor, 12Gb of RAM which 10Gb were given to heap size memory of Weka and a Windows 10 operating system.

The objective was to apply the algorithms and see what performed better accordingly to its characteristics. Then, compare if the performance, with the data and approach of this study in data pre-processing and data transformation, was higher than the best algorithms of each article referred in Literature Review chapter. Where Machine Learning algorithms were used for a binary prediction of flight delay existence on an individual airport similarly to this work.

Through the results achieved (when achieved because of ”run out of memory” errors of Weka) in the phase of implementation, it was necessary to weight the pros and cons of the metrics chosen more specifically the ROC Curve, F-Measure, and Accuracy when selecting the best approaches.

Although the problem of the unbalanced dataset was treated via undersampling and oversampling, the accuracy was a metric that had less weight in the decision. It was used only when small differences were detected in the other measures, to minimize the minimal hypothesis that the algorithms still preferred the majority class over the minority one.

For analysis, results are presented from four different datasets for each approach of sampling technique, SMOTE or undersampling. The first two datasets relate to the ones that include variables that just provide the prediction of the delay after the flight takes off such as “Real_Dep_Time” and “Dep_Delay”. Thus, the first includes the existence of outliers and the second does not include them.

The last two datasets, contrarily to the first two, do not account for the two variables. This makes possible for the prediction of the delay in the beginning of the day of the flight where the first of the last two encompass the existence of possible outliers and the other does not. The possibility for only predict the delay in the beginning of the day is because of the existence of other variables that needs information of the previous day.

As a general analysis of Annex 11 and 12 of the Annexes chapter, and in contrast to what was expected, it can be seen that the SMOTE technique has better results in terms of ROC Curve than the undersampling technique, similarly to what Batista, Prati and Monard (2004) concluded.

It is also observed in the two annexes mentioned above, that the results of the datasets comprising the two variables, real time of departure and delay in departure, have higher values in the ROC Curve as expected. However, it has the inconvenience of the prediction being made only after the flight departs. The non-accounting of this type of variables can decrease in average 0,3 decimals in the ROC Curve metric but in some cases, can increase the correctly classified instances metric.

In addition, is possible to see that when SMOTE is applied there is a higher accuracy and ROC Curve while outliers are not considered. Meanwhile, when undersampling is used this difference is not

noticed, being slightly equal and in some cases the existence of outliers makes the model perform better.

The influence of the outliers in helping to a better model can be caused because of the high probability of extreme cases in this type of issue. For example, the extreme values of some weather variables that, in the USA, are probable to happen.

Nonetheless, when looking at the datasets that does not account for the two variables it is witnessed the best performance in terms of the ROC Curve in the dataset where the outliers are considered for both techniques, SMOTE and Undersampling.

By the analysis of the all the results (Annex 11 and 12 of the Annexes chapter), the best attribute selection approach for each algorithm (Decision Tree J48, Random Forests and Multilayer Perceptron) of a specific technique (SMOTE and Undersampling) is selected respectively to each dataset (with or without variables of departure information and with or without outliers' removal) as it can be seen in the table below.

		1: Flight Dataset + Dep_Delay & Real_Dep_Time + Outliers			*2: Flight Dataset + Dep_Delay & Real_Dep_Time - Outliers*			*3: Flight Dataset - Dep_Delay & Real_Dep_Time + Outliers*			*4: Flight Dataset - Dep_Delay & Real_Dep_Time - Outliers*		
		ROC	*F*	*CCI*	*ROC*	*F*	*CCI*	*ROC*	*F*	*CCI*	*ROC*	*F*	*CCI*
SMOTE	*J48*	0.84	0.82	79.77	0.83	0.82	79.77	0.53	0.80	83.06	0.50	0.75	74.88
		GainRatio (10)			CfsSubsetEval (4)			GainRatio (10)			GainRatio (10)		
	RF	0.83	0.82	79.76	0.83	0.82	80.00	0.53	0.78	80.89	0.52	0.79	84.26
		GainRatio (15)			GainRatio (10)			GainRatio (10)			GainRatio (10)		
	MLP	0.89	0.82	79.72	0.89	0.86	84.06	0.56	0.79	85.63	0.51	0.79	85.67
		GainRatio (15)			GainRatio (10)			GainRatio (15)			GainRatio (10)		
Undersampling	*J48*	0.83	0.82	79.76	0.83	0.82	79.77	0.57	0.73	69.02	0.50	0.79	85.67
		GainRatio (10)			GainRatio (20)			GainRatio (15)			CfsSubsetEval (6)		
	RF	0.85	0.82	79.62	0.83	0.78	74.88	0.53	0.56	48.67	0.52	0.61	54.51
		GainRatio (10)			GainRatio (15)			CfsSubsetEval			CfsSubsetEval (6)		
	MLP	0.89	0.82	79.76	0.84	0.82	79.77	0.50	0.79	85.67	-	-	-
		CfsSubsetEval (3)			CfsSubsetEval (3)			GainRatio (10)			-		

Table 9: Resume of the best results for each technique, algorithm, and dataset
Source: Made by the author

A. Dataset 1: Flight Dataset + Dep_Delay & Real_Dep_Time + Outliers

By the use of SMOTE oversampling technique, the models chosen for J48, RF and MLP were the ones with the higher ROC Curve. In the case of the RF algorithm, the one chosen was not the one with the highest accuracy nonetheless, having a minimal difference to the model selected.

For the Undersampling technique only the models chosen for J48 and RF were the ones with the highest ROC Curve value. On the other hand, the MLP algorithm preferred was not the one with the higher ROC Curve but instead the one with the second highest value. This was because the accuracy was well depreciated comparing to the one chosen (accuracy of 28.17 with a ROC Curve of 0.89 and F-Measure of 0.28 compared to the one chosen with 79.76 of acuracy, ROC Curve of 0.89 and F-Measure of 0.82).

B. Dataset 2: Flight Dataset + Dep_Delay & Real_Dep_Time - Outliers

Concerning SMOTE technique the models chosen for J48 and MLP were the ones with the higher ROC Curve, F-Measure and Accuracy values. Altought, in the case of the RF algorithm, the model chosen was not the one with the higher ROC Curve neither the one with the higher accuracy. Instead, it was a result that represented a compromise between the two models. Thus, the model chosen has a higher accuracy then the one with a higher ROC Curve (79.77 of accuracy and 0.83 of ROC Curve comparing to one selected with a 80.00 accuracy and 0.83 ROC Curve) and a higher ROC Curve than the one with the higher accuracy (80.12 of accuracy and 0.80 of ROC Curve comparing to the chosen one).

About the Undersampling technique only the models selected for MLP and RF were the ones with the higher ROC Curve, F-Measure and Accuracy. Regarding the J48 algorithm, all different attribute selection techniques had the same values in the metrics, and for that reason, we resort to the time metric. This metric made choose the approach that selected twenty attributes having a minimal difference between selecting 10 or 20 features (1.16 seconds to build the model and 0.28 seconds for testing it comparing to the values of the chosen approach with 1.18 for building and 0.31 for testing). This choice was also because its preferable to have more variables than less when having the same results and little time difference between build and testing the model.

C. Dataset 3: Flight Dataset - Dep_Delay & Real_Dep_Time + Outliers

About the SMOTE technique results is important to see the decisions made for each of one of the algorithms. In the case of the J48 algorithm, the approach selected was the one with the higher accuracy and F-Measure but not the higher ROC Curve value (83.06 accuracy, 0.80 of F-Measure and 0.53 ROC Curve value comparing to the one with the higher ROC Curve with 80.49 of accuracy, 0.79 of F-Measure and 0.54 of ROC Curve value). This decision was due to the one with the higher ROC Curve had three more variables than the one chosen, not justifying the high decrease of the accuracy and negligible increase of the ROC Curve value.

For the RF algorithm, the model chosen was the one with higher accuracy but not the higher ROC Curve value (76.33 of accuracy, 0.77 of F-Measure and the highest ROC Curve value of 0.55, comparing to the one chosen with an accuracy value of 80.89, 0.78 of F-Measure and 0.53 of ROC Curve, with an increase of, approximately, 5 percentage points of accuracy and a minimal decrease of 0.02 between ROC Curves).

In reference to the MLP algorithm , the preferred model was the one that had the higher accuracy. The reason was because, from the chosen one to the one with the highest ROC Curve, it is noticed that there is a higher gain of accuracy (approximately, more 2 percentage point) than a loss of ROC Curve (decrease of 0.02) and also with a slightly higher F-Measure than the one in comparison.

For the Undersample tecnhique, the models chosen for RF and MLP were the ones with the highest ROC Curve and Accuracy values. In the case of J48, two attribute selection approach had the highest ROC Curve and Accuracy values but preferring the approach that had more variables with a minimal increase of runtime (13.42 seconds for building the model and 0.54 for testing, comparing to the approach chosen with 14.47 seconds for building the model and 0.25 for testing).

D. Dataset 4: Flight Dataset - Dep_Delay & Real_Dep_Time - Outliers

Regarding the SMOTE oversampling technique, the models distincted for the J48 and RF were the ones with the higher ROC Curve values. In relation to the model chosen to the MLP was the one with the highest accuracy. Due to a high decrease of the accuracy and F-Measure and to a smaller one in the value of the ROC Curve (85.67 of accuracy, 0.79 of F-Measure and 0.51 of ROC Curve comparing to 56.54 of accuracy with a decrease of approximately 29 percentage points, 0.63 of F-Measure with a decrease of 0.16 and 0.53 of ROC Curve that only increases in 0.02) when comparing to the one with the highest ROC Curve value.

For the Undersampling technique, the model selected for the RF algorithm was the one with the higher ROC Curve and Accuracy values. In respect of the J48 algorithm, the model preferred was the one where is presence a higher Accuracy and F-Measure instead of the one with the highest ROC Curve (85.67 of accuracy, 0.79 of F-Measure and 0.50 of ROC Curve against the values of 71.77 of accuracy, 0.74 of F-Measure and 0.56 of ROC Curve value, respectively) and also a much faster performance.

In the case of the MLP the results for the approaches were impossible to achieved due to the the issue of "run out of memory" error, caused by the inclusion of variables with high number of categorical classes which is the case of the variables "*Flight_Num*" and "*Mod*".

This happened because, with a smaller dataset caused by the application of the undersampling technique, the categorical variables were more recurrent used than within the other technique, where numeric variables were used more times as it can be seen in Figure 42.

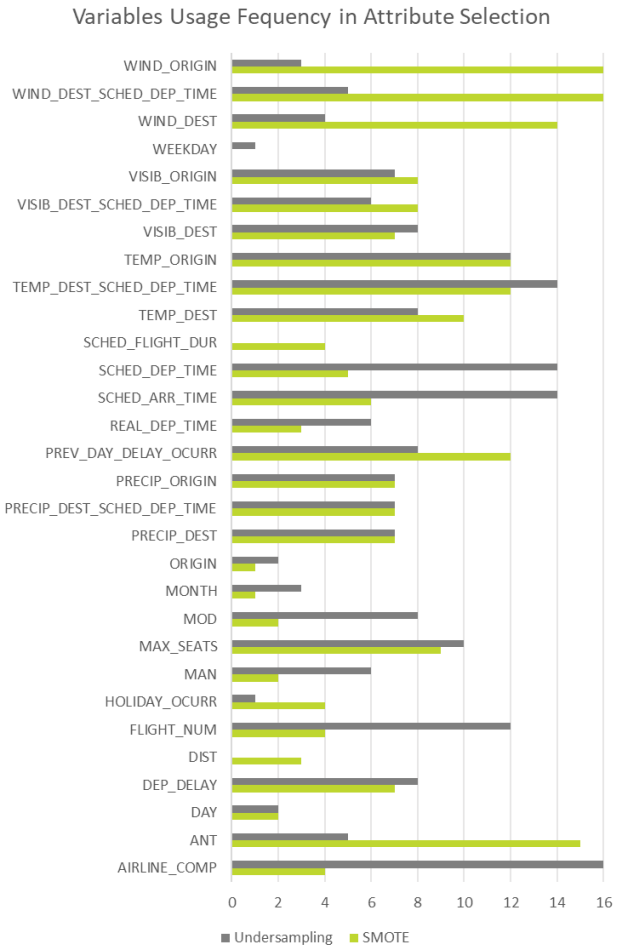


Figure 42: Variables Usage Frequency by Sampling Technique
Source: Made by the author

It can be seen, for example, that in SMOTE the variables that are selected more frequently through the various approaches of attribute selection (CfsSubsetEval, GainRatio(10), GainRatio(15) and GainRatio(20)) are the ones of numerical type. When it comes to undersampling, despite having numerical variables frequently in use, there is a higher presence of categorical variables chosen more times. For example, it can be seen that the variable "Airline_Comp", always selected through undersampling (sixteen times out of the sixteen times, four attribute selection approach for each four datasets), it is only selected four times in the SMOTE technique. The same can be seen through the variable "Flight_Num", the one that compromises the implementation of MLP in undersampling

because of being selected in twelve times of the sixteen, and in SMOTE only being selected in four times and impairing few cases.

After the selection described above, and from the comparison between the use of SMOTE or undersampling, the best approaches were selected from the ones already selected. This was based on the metrics that had higher results, as it can be seen in the table below and where it can be seen that most of the datasets had higher results in each algorithm by the application of the SMOTE technique.

	1: Flight Dataset + Dep_Delay & Real_Dep_Time + Outliers			*2: Flight Dataset + Dep_Delay & Real_Dep_Time - Outliers*			*3: Flight Dataset - Dep_Delay & Real_Dep_Time + Outliers*			*4: Flight Dataset - Dep_Delay & Real_Dep_Time - Outliers*		
	ROC	*F*	*CCI*	*ROC*	*F*	*CCI*	*ROC*	*F*	*CCI*	*ROC*	*F*	*CCI*
J48	0.84	0.82	79.77	0.83	0.82	79.77	0.53	0.80	83.06	0.50	0.79	85.67
	SMOTE			SMOTE			SMOTE			Undersampling		
RF	0.85	0.82	79.62	0.83	0.82	80.00	0.53	0.78	80.89	0.52	0.79	84.26
	Undersampling			SMOTE			SMOTE			SMOTE		
MLP	0.89	0.82	79.72	0.89	0.86	84.06	0.56	0.79	85.63	0.51	0.79	85.67
	SMOTE			SMOTE			SMOTE			SMOTE		

Table 10: Resume of the best results for each algorithm and dataset
Source: Made by the author

Consequently, it is needed to see which approach performs better, either by including the outliers or removing them. For that reason, a selection was made to the table above by selecting the best approaches for each algorithm when the inclusion of the variable of Dep_Delay and Real_Dep_Time is considered and when it is not. As a result of this step, the table below was constructed.

	Including Dep_Delay and Real_Dep_Time						*Excluding Dep_Delay and Real_Dep_Time*					
	Training Set			*Test Set*			*Training Set*			*Test Set*		
	ROC	*F*	*CCI*	*ROC*	*F*	*CCI*	*ROC*	*F*	*CCI*	*ROC*	*F*	*CCI*
J48	0.95	0.88	94.62	0.84	0.82	79.77	0.81	0.86	87.48	0.53	0.80	83.06
	With Outliers						With Outliers					
	SMOTE						SMOTE					
	GainRatio (10)						GainRatio (10)					
RF	1	0.99	99.14	0.85	0.82	79.62	1	1	100	0.52	0.79	84.26
	With Outliers						Without Outliers					
	Undersampling						SMOTE					
	GainRatio (10)						GainRatio (10)					
MLP	0.88	0.93	94.01	0.89	0.86	84.06	0.73	0.75	78.06	0.56	0.79	85.63
	Without Outliers						With Outliers					
	SMOTE						SMOTE					
	GainRatio (10)						GainRatio (15)					

Table 11: Resume of the best results for each algorithm and corresponding training results by two views: including Dep_Delay and Real_Dep_Time and otherwise
Source: Made by the author

It is possible to observe that most of the times all classifiers performed better on the training data then when predicting with unseen data. This happens with the exception of the MLP algorithm when excluding the two variables. This is probably because they could overfit modeling in every minor variation of the input data despite the attempt to eliminate this problem.

Nonetheless, only in the classifiers where the variables of "*Dep_Delay*" and "*Real_Dep_Delay*" are excluded the values of ROC Curve most differentiate. Despite the fact that this difference also exists in the models where the two variables are considered, it can be seen that is not so drastic comparing to the previous situation.

It is noticeable, with the help of Table 11, a decrease of approximately between 0.2 and 0.3 in the ROC Curve when not using the variables.

However, if we want to predict if a flight will be late at the arrival before the travel, it is obvious that this type of variables should not be included. This will give to the airlines and airports a margin for taking actions or improve their services at the arrival airport. For that purpose, the algorithm that is most suitable, although with a lower ROC Curve value, is the MLP using the SMOTE technique, including the outliers and selecting the attribute through the GainRatio approach retaining ten variables.

If the interest of the study is to predict if a flight is going to suffer from delay at the arrival at the destination after taking off, then the variables can be taking into account. For this purpose, the most suitable algorithm is the MLP using the SMOTE technique but excluding the outliers and with the selection of the variables through the GainRatio selecting ten variables.

Furthermore, it is visible by the observation of the table above that the decision trees have a smaller performance in the ROC Curve value than the RF and MLP when considering the two variables. And a higher ROC Curve value than RF but lower than MLP when not considering the two variables. We conclude that by its simplicity, understandability and capacity to handle categorical variables contrarily to the others (as it can be seen in Annex 11 and 12 of the Annexes chapter) they still do not have higher performances.

When looking to the variables that most contribute to the delay in the best algorithms selected (presented in Table 11) we can see with the help of the images below that the most important variables are firstly the ones related to the weather being selected in all the models.

Secondly, the ones related to the information of the characteristics of the airplane such as the antiquity and the number of maximum seat, used in five of the models. Then, variables of propagation of the delay such as the occurrence of delay in the previous day at the origin is used in three models selected. At last and used in one of the models, the variables of flight information like airline company and flight number.

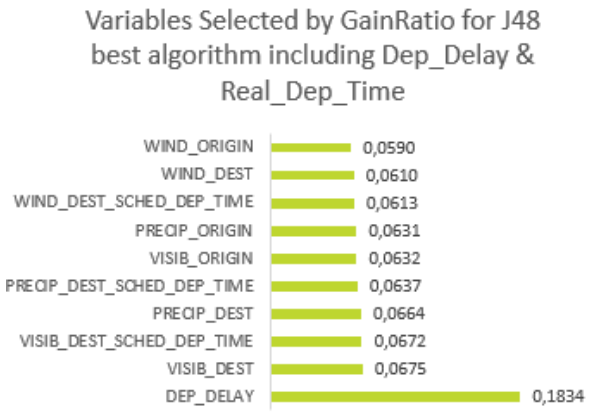


Figure 43: Variables Selected by GainRatio for the J48 best algorithm including Dep_Delay & Real_Dep_Time variables
Source: Made by the author

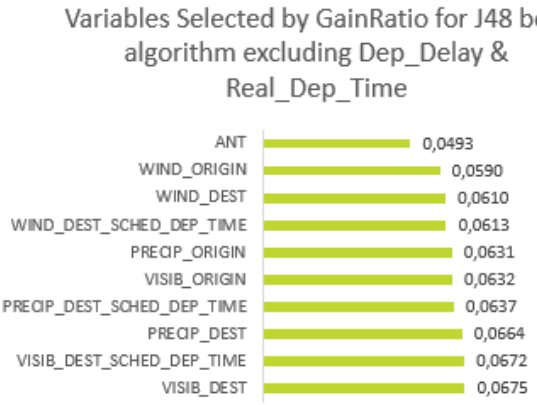


Figure 44: Variables Selected by GainRatio for the J48 best algorithm excluding Dep_Delay & Real_Dep_Time variables
Source: Made by the author

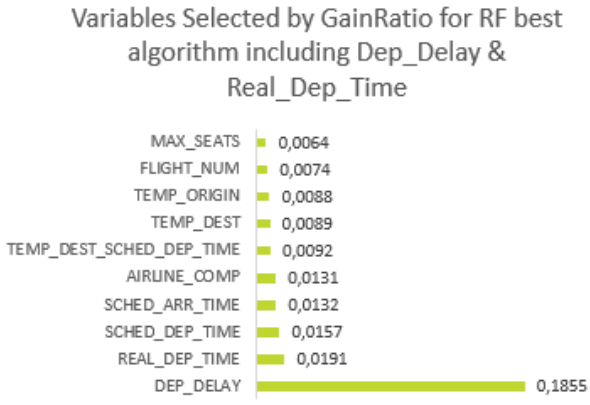


Figure 45: Variables Selected by GainRatio for the RF best algorithm including Dep_Delay & Real_Dep_Time variables
Source: Made by the author

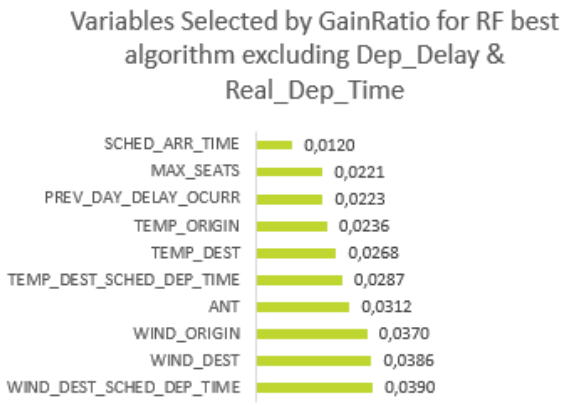


Figure 46: Variables Selected by GainRatio for the RF best algorithm excluding Dep_Delay & Real_Dep_Time variables
Source: Made by the author

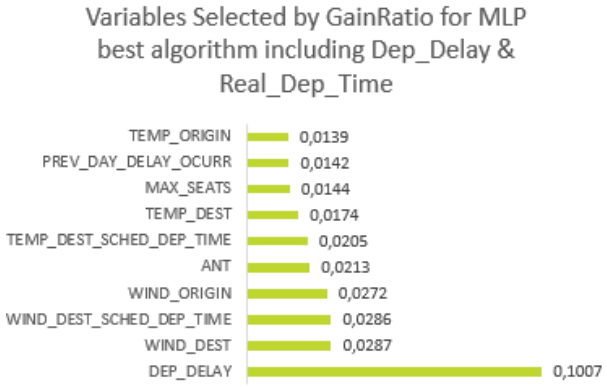


Figure 47: Variables Selected by GainRatio for the MLP best algorithm including Dep_Delay & Real_Dep_Time variables
Source: Made by the author

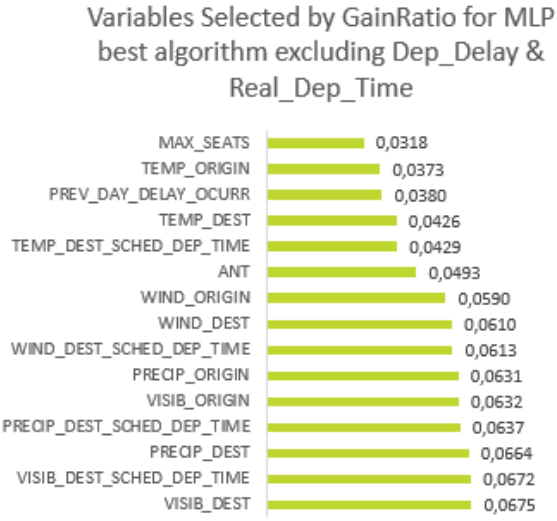


Figure 48: Variables Selected by GainRatio for the MLP best algorithm excluding Dep_Delay & Real_Dep_Time variables
Source: Made by the author

It is important to note that when the variable of Dep_Delay is included in the model is always the one with the highest rank value. This is because of what was already mentioned and also concluded by M. S. Kim (2016), that the inclusion of this type of variables gives to the model a huge information when modeling the delay in the arrival, being a variable of great importance when included.

When comparing the results of the algorithms of this study with the other studies, that also studied the delay at individual airports using machine learning algorithms (Belcastro et al., 2016; Choi et al., 2016; Y. J. Kim et al., 2016), it is possible to distinguish the approaches and results obtained by them, and by this work as present in Table 12.

It is important to note that these three studies differ from this by the use of different variables and sources. By comprising different periods of time. And also, by applying different approaches such as in article of Choi et al. (2016) where they employed 10-Fold Cross Validation, transformed all categorical variables to numerical and pre-processed the data gathered to test the model in the same way that pre-processed the training set.

In these cases, the variables of Dep_Delay and Real_Dep_Time were not used in the model. For that reason, the results used to compare this work with the others are also referring to the results where these two variables were not considered.

Without knowing information about the area under the ROC Curve for the other studies test results, but knowing that all the datasets are balanced, including this work, each of the above articles can be analyzed with this study.

It is interesting to note that, in reference to the MLP algorithm, the other authors have accomplished a higher accuracy than the one attained in this study (87.42% compared to 85.63%). Regarding the RF algorithm the results here obtained are much better than the ones from the other studies. Furthermore, although there is no information of ROC Curve in the test results of the other articles we know that the value accessed in this work is low. Nevertheless, when comparing the ROC Curve of the training data of the article of Choi et al. (2016) we can denote that the one computed by the model in this study is higher (0.68 comparing to 1 (presented in Table 11), respectively).

Article	*Information used*	*Objective*	*Type of Categorical Prediction*	*Algorithms Accuracy*	
				Random Forest	*Multilayer Perceptron*
(Y. J. Kim et al., 2016)	Delay status of the day; Historical flight data; Weather data	Predict Class of Delay of an individual flight	Multiclass	-	87.42%
(Belcastro et al., 2016)	Historical Flight data; Weather data	Predict arrival delays of individual flight due to weather conditions	Binary	74.20%	-
(Choi et al., 2016)	Historical Flight data; Weather data	Predict arrival delays of individual flight	Binary	80.36%	-
Our work	Historical Flight data; Weather data; Airplane info; Delay Propagation information	Predict arrival delays of individual flight	Binary	84.26%	85.63%

Table 12: Related Work Comparison
Source: Made by the author, based on (Belcastro et al., 2016)

Additionally, when one compares the present work results to the sites that predict if a flight is going to be delayed or not one can see that, the accuracy performance of their models is higher than 80% which could be compared to this work's accuracy performance. This comparison could not be done directly because of the different information variables and because of the higher forecast horizon.

However, by the same conditions of not considering the variables of Dep_Delay and Real_Dep_Time to have a prediction before to the time of the flight, this work's best approach that is the MLP algorithm has as accuracy a value of 85.63%.

It stands in the same level of the DelayCast performance. Higher than FlighCaster, where data is not balanced and having a recall of 60% compared to the result of this study with a recall of 85.6%. Nonetheless, lower than the KnowDelay that can predict a flight to be delayed due to weather over three days in advance but also having high results by overpredicting bad weather.

Platform	*Information used*	*Performance*
FlightCaster	Flight; Weather	Accuracy: 85% Recall: 60% (Smartertravel, 2009)
KnowDelay	Airport Performance; Weather	Accuracy: 90% (Knowdelay, 2018)
DelayCast	Flight; Weather; Airline Company History; Number of Passengers	Accuracy: 80-90% (Tourism Review, 2008)

Table 13: Related Platforms Comparison
Source: Made by the author, based on (Belcastro et al., 2016)

5. CONCLUSIONS

Knowledge is power.

– Francis Bacon (1561-1626)[6]

This project has as its main objective the prediction of the occurrence of a delay, in the arrivals of Hartsfield-Jackson International Airport, being able to know which variables contribute the most to the existence of delay.

For reaching this objective the following steps had to be achieved:

- Construct a database with information concerning the flights and additional information;
- Explore the delays accordingly to different variables;
- Construct a predictive model using Data Mining and Machine Learning techniques to predict the delay of a flight in the arrival;
- Apply the model developed in new data to make predictions and see which fits better to the problem according to the desired advance of the prediction;
- Identify the variables that contribute most to the existence of delay.

The first step was accomplished by the construction of a dataset to present to the algorithm via the use of the Excel and SQL Management Studio. It contemplated information of flights and airplanes but also weather, time zone, flight durations and holidays.

The second step was possible through the use of the dataset constructed, and accordingly to the observations gathered. It was possible to note that the months with a higher percentage of delays were the ones occurring in the Winter and Summer seasons. Higher distances have a higher percentage of delayed flights being seen that most of flights are of short distances. The schedules departures and arrival times have a higher percentage of delays in the afternoon and evening. Airports of origin located in the south of the USA are the ones with the higher percentage of delay. Is interesting that, when the airplane operating the flight has a higher number of seats (>300), the percentage of delayed flights increases. When looking for the role of the flights that are delayed or canceled in the previous day of a flight at its origin, it is seen that when the number increases the percentage of delayed flights also increases. As a confirmation of the already known by intuition, extreme values in temperature increases the percentage of delayed flights. Higher levels of precipitation have fewer numbers of flights, but the ones that happen has a high percentage of delayed flights. Higher velocity of the wind causes an increase of the delayed flight percentage. And, a loss of visibility increases the percentage of delayed flights.

With the insights obtained by the visualization of the data, the preprocessing and transformation was carried out for the building of a model capable of predicting the delay information concluding the third step of the specific objectives.

[6] Francis Bacon, 1st Viscount St Alban, was an English philosopher, statesman, scientist, jurist, orator, and author (for more information consult https://en.wikipedia.org/wiki/Francis_Bacon)

For performing the fourth step, the model obtained was applied to new data to see how well it could perform when predicting in the unknown data.

Concerning the different advance of the prediction that we could have, by including or excluding the variables of Dep_Delay and Real_Dep_Time, it was possible to achieve better performance with the MLP algorithm among the others.

When considering the two variables, the MLP performed better by the use of SMOTE technique, without the inclusion of the outliers and including the ten variables most important accordingly to the GainRatio. If not considering the two variables then the MLP algorithm performs better by also using the SMOTE technique, with the inclusion of the outliers and the fifteen variables most important also accordingly to the GainRatio.

By the conclusions made in the Results and Discussion chapter, and responding to the fifth step, it can be seen that variables of weather, in general, were the ones with higher importance in the best model of each advance of the prediction. When the variable "*Dep_Delay*" is included is the one that had higher gain ratio because of the access to this information improves the predictions as seen in the Results and Discussion chapter.

Also, the variables "*Ant*" and "*Max_seats*" are considered in both models but with less importance, happening the same with the variable created about the amount of delayed and canceled flights in the previous day at the origin airport ("*Prev_Day_Delay_Occur*").

Contrarily to what was expected, variables such as the day or month, scheduled departure and arrival time, distances, and other variables turn out to not be such good than the ones chosen in these two models.

6. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

The enemy of art is the absence of limitation.

– Orson Welles (1915-1985)[7]

Recognizing the limitations of a study is an opportunity to improve and make suggestions for further research, hence in this chapter, the limitations that restricted the development of this study are present. Ideas that may be developed in future works, majorly resulting from limitations, according to this theme are also presented.

The first limitation that influences all the course of this study is concerning to the availability of online repositories where data about several issues are accessible to all. Originally the aim was to conduct this study about Portugal which contrarily to the USA does not have, or is not easily accessible, a repository for federal government information – data.gov. In this type of repository is possible to have public access to high value data, tools, and resources to conduct researches, or even the availability of data, in this case about transport, on the website of the Department of Transportation. Furthermore, attempts to contact ANA Airports for reaching possible data for the Lisbon airport were made with no success. This limitation leads to the first recommendation, an implementation of this work with data concerning to Portugal, more specifically about one of the existent airports.

The following limitation can be related to the tool, for implementing all steps, practically and simply. Initially, the use of SAS Enterprise Miner was the chosen for developing this project. While doing the review of the methodology desired to implement, we come across with the complexity and the knowledge needed for the process of implementing SMOTE technique in this software. As a solution, it was adopted the Weka Software with a large availability of machine learning techniques and algorithms.

The task of dealing with a large dataset, made the Weka software, in some situations, block or run out of memory. The solution was to assign a higher memory to Weka being given 10 out of 12Gb of the RAM available by the computer where this project was implemented. Nonetheless, in some situations, it was not possible of doing what was intended because of running out of memory error. Such situations happened when it was wanted to see if the algorithm performed better without an attribute selection than with it and, for that reason, this was not possible concluding that, for higher datasets, Weka probably is not a feasible tool.

The circumstance mentioned above, lead to another recommendation for future works when dealing with smaller datasets. To apply attribute selection techniques that are very computing slow, but which can improve the results, such as the Wrapper method or Embedded method. It is also interesting to extend this study to other algorithms that were not taking into consideration in the present work. Also, if a huge dataset is in hands, perhaps an approach to other techniques like Python or Apache Hadoop could be advantageous when the existent of knowledge about them.

[7] George Orson Welles was an American actor, director, writer, and producer who worked in theatre, radio, and film (for more information consult https://en.wikipedia.org/wiki/Orson_Welles)

Another extent to future work, which could contribute to airline and airports improvement of their behavior when dealing with this phenomenon is to know the number of minutes inherent to an existent delay by performing a regression problem.

A further issue is about the tuning of the algorithms. The approach followed was to adopt the default settings already defined by Weka software. However, with a little more experience and time, it would be interesting to test whether, by changing the parameters, better results would be obtained.

Additionally, it is important to mention the essential role of the variables selected for explaining this theme that, sometimes, is an important factor in model performance. In this specific case, despite the effort made to acquire variables that, in the time, seemed to be interesting for the model, it was noted that, through variable selection, a large number of variables were not important when explaining the model. Consequently, future work on variables transformation and selection of other features related to flights could be a meaningful step towards a model with higher performance.

Finally, it is proposed as future work, a predictive model that can overcome all the lacks here existent and fulfill the needs not addressed.

7. REFERENCES

Abdel-Aty, M., Lee, C., Bai, Y., Li, X., & Michalak, M. (2007). Detecting periodic patterns of arrival delay. *Journal of Air Transport Management*, *13*(6), 355–361. https://doi.org/10.1016/j.jairtraman.2007.06.002

AhmadBeygi, S., Cohn, A., Guan, Y., & Belobaba, P. (2008). Analysis of the potential for delay propagation in passenger airline networks. *Journal of Air Transport Management*, *14*(5), 221–236. https://doi.org/10.1016/j.jairtraman.2008.04.010

Airports Council International. (2016). *Preliminary World Airport Traffic Ranking*. Retrieved from http://www.aci.aero/News/Releases/Most-Recent/2016/04/04/ACI-releases-preliminary-world-airport-traffic-rankings-

Allison, P. D. (2001). Missing Data. In *Quantitative Applications in the Social Sciences* (pp. 72–89). https://doi.org/10.1136/bmj.38977.682025.2C

Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. *IADIS European Conference Data Mining*, (January), 182–185. Retrieved from http://recipp.ipp.pt/handle/10400.22/136

Balakrishna, P., Ganesan, R., Sherry, L., & Levy, B. S. (2008). Estimating Taxi-out Times with a Reinforcement Learning algorithm. *AIAA/IEEE Digital Avionics Systems Conference - Proceedings*, 1–12. https://doi.org/10.1109/DASC.2008.4702812

Ball, M., Barnhart, C., Dresner, M., Neels, K., Odoni, A., Peterson, E., … Hansen, M. (2010). *Total Delay Impact Study -- A comprehensive assessment of the cost and impacts of flight delay in the United States*. Retrieved from http://www.isr.umd.edu/NEXTOR/pubs/TDI_Report_Final_10_18_10_V3.pdf

Basheer, I. A., & Hajmeer, M. (2000). Artificial neural networks: Fundamentals, computing, design, and application. *Journal of Microbiological Methods*, *43*(1), 3–31. https://doi.org/10.1016/S0167-7012(00)00201-3

Batista, G. E. A. P. A., & Monard, M. C. (2002). Proceedings of the First International Workshop on Data Cleaning and Preprocessing. In S. Zhang, Q. Yang, & C. Zhang (Eds.), *An Analysis of Four Missing Data Treatment Methods for Supervised Learning* (pp. 142–152). Maebashi, Japan. Retrieved from https://www.researchgate.net/profile/Stefano_Cerri/publication/232866130_GAsRULE_for_Knowledge_Discovery/links/09e4150c6f1905948f000000/GAsRULE-for-Knowledge-Discovery.pdf

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. *ACM SIGKDD Explorations Newsletter*, *6*(1), 20–29. https://doi.org/10.1145/1007730.1007735

Belcastro, L., Marozzo, F., Talia, D., & Trunfio, P. (2016). Using Scalable Data Mining for Predicting Flight Delays. *ACM Transactions on Intelligent Systems and Technology*, *8*(1), 1–20. https://doi.org/10.1145/2888402

Belobaba, P., Odoni, A., & Barnhart, C. (2009). *The global airline industry*.

Bureau of Transportation Statistics. (2016a). Airline On-Time Performance and Causes of Flight Delays. Retrieved October 31, 2016, from http://www.rita.dot.gov/bts/help_with_data/aviation/index.html#q8

Bureau of Transportation Statistics. (2016b). On-time Performance. Retrieved October 31, 2016, from http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236

Bureau of Transportation Statistics. (2017). On-Time Performance - Flight Delays at a Glance. Retrieved July 17, 2017, from https://www.transtats.bts.gov/HomeDrillChart.asp?URL_SelectMonth=4&URL_SelectYear=2017

Chawla, N. V. (2009). Data Mining for Imbalanced Datasets: An Overview. In *Data Mining and Knowledge Discovery Handbook* (pp. 875–886). https://doi.org/10.1007/978-0-387-09823-4_45

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Chen, P. P.-S. (1976). The Entity-Relationship Model - Toward a Unified View of Data. *ACM Transactions on Database Systems*, *1*(1), 9–36. https://doi.org/10.1145/320434.320440

Choi, S., Kim, Y. J., Briceno, S., & Mavris, D. (2016). Prediction of Weather-induced Airline Delays Based on Machine Learning Algorithms. *AIAA/IEEE Digital Avionics Systems Conference - Proceedings*, *2016–Decem*, 1–6. https://doi.org/10.1109/DASC.2016.7777956

Dash, M., & Liu, H. (1997). Feature selection for classification. *Intelligent Data Analysis*, *1*(3), 131–156. https://doi.org/10.3233/IDA-1997-1302

Dean, J. (2014). *Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners*. New Jersey: John Wiley & Sons, Inc.

De Ville, B. (2001). *Microsoft Data Mining: Integrated Business Intelligence for E-Commerce and Knowledge Management*. Digitalm Press. Retrieved from https://books.google.pt/books?id=iFY7ovqjUhgC&pg=PA94&lpg=PA94&dq=Data+preparation+is+60%25+of+effort+for+data+mining+process+(Pyle)&source=bl&ots=FtS26-5DfB&sig=SfnEmmm_Kc1ClcNHe3REU67R_E0&hl=pt-PT&sa=X&ved=0ahUKEwj3msG29I7XAhWGshQKHa-TBtMQ6AEIPjAD#v=one

Department of Agronomy. (2017a). Iowa State University, Iowa Environmental Mesonet. Retrieved November 14, 2017, from https://mesonet.agron.iastate.edu/info/iem.php

Department of Agronomy. (2017b). Iowa State University, Iowa Environmental Mesonet. Retrieved November 14, 2017, from https://mesonet.agron.iastate.edu/request/download.phtml

Department of Agronomy. (2017c). Iowa State University, Iowa Environmental Mesonet. Retrieved November 14, 2017, from https://mesonet.agron.iastate.edu/ASOS/

Fahrner, C., & Vossen, G. (1995). A survey of database design transformations based on the Entity-Relationship model. *Data and Knowledge Engineering*, *15*(3), 213–250. https://doi.org/10.1016/0169-023X(95)00006-E

Fayyad, U. (n.d.). Knowledge Discovery in Databases: An Overview, 5–16.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, *17*(3), 37. https://doi.org/10.1609/aimag.v17i3.1230

Federal Aviation Administration. (2015). Forming an N-Number. Retrieved August 12, 2017, from https://www.faa.gov/licenses_certificates/aircraft_certification/aircraft_registry/forming_nnu

mber/

Federal Aviation Administration. (2016a). Aircraft Registry. Retrieved October 31, 2016, from http://www.faa.gov/licenses_certificates/aircraft_certification/aircraft_registry/

Federal Aviation Administration. (2016b). FAA Registry: Make/Model Inquiry. Retrieved December 20, 2016, from http://registry.faa.gov/aircraftinquiry/acftref_inquiry.aspx

Federal Aviation Administration. (2016c). FAA Registry: N-Number Inquiry. Retrieved December 20, 2016, from http://registry.faa.gov/aircraftinquiry/NNum_inquiry.aspx

GE Aviation. (2012). GE Flight Quest: Think you can change the future of flight? Retrieved July 20, 2017, from https://www.kaggle.com/c/flight

Grubbs, F. E. (1974). *Procedures for Detecting Outlying Observations in Samples*. Retrieved from http://www.dtic.mil/dtic/tr/fulltext/u2/781499.pdf

Guimerà, R., & Amaral, L. A. N. (2004). Modeling the world-wide airport network. *European Physical Journal B*, *38*(2), 381–385. https://doi.org/10.1140/epjb/e2004-00131-0

Hall, M. A., & Holmes, G. (2002). *Benchmarking Attribute Selection Techniques for Discrete Class Data Mining*. *IEEE Transactions on Knowledge and Data Engineering* (Vol. 15). Hamilton, New Zealand. https://doi.org/10.1109/TKDE.2003.1245283

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *Sigkdd Explorations*, *11*(1), 10–18.

Hamon, J. (2013). *Optimisation combinatoire pour la sélection de variables en régression en grande dimension: Application en génétique animale*. Université des Sciences et Technologie de Lille. Retrieved from http://hal.inria.fr/tel-00920205/

Han, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques*. *Elsevier* (2nd ed., Vol. 12). Morgan Kaufmann Publishers, Inc. https://doi.org/10.1007/978-3-642-19721-5

Hand, D., Hand, D., Mannila, H., Mannila, H., Smyth, P., & Smyth, P. (2001). *Principles of data mining*. *Drug safety : an international journal of medical toxicology and drug experience* (Vol. 30). https://doi.org/10.2165/00002018-200730070-00010

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction Second Edition* (2nd ed.). Springer. https://doi.org/10.1007/b94608

Hawkins, D. M. (1980). Introduction. In *Identification of Outliers* (pp. 1–12). Springer, Dordrecht. https://doi.org/10.1007/978-1-4614-0406-4

Hernández, M. A., & Stolfo, S. J. (1998). Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem. *Data Mining and Knowledge Discovery*, *2*(January), 9–37.

Herzmann, D., Klein, W., & Taylor, E. (2013). Iowa State University, Extension and Outreach. Retrieved November 14, 2017, from https://www.extension.iastate.edu/article/iowa-environmental-mesonet-data-used-thousands-every-day

Hoens, T. R., & Chawla, N. V. (2013). Imbalanced Datasets: From Sampling to Classifiers. In H. He & Y. Ma (Eds.), *Imbalanced Learning: Foundations, Algorithms, and Applications, First Edition.* (1st Editio, pp. 43–59). https://doi.org/10.1002/9781118646106.ch3

International Air Transport Association (IATA). (2015). *Air Passenger Market Analysis*. Retrieved from

http://www.iata.org/whatwedo/Documents/economics/passenger-analysis-dec-2015.pdf

International Civil Aviation Organization. (2011). *Manual of Aeronautical Meteorological Practice*. (International Civil Aviation Organization, Ed.) (9th ed.).

Ionescu, L., Gwiggner, C., & Kliewer, N. (2016). Data Analysis of Delays in Airline Networks. *Business and Information Systems Engineering*, *58*(2), 119–133. https://doi.org/10.1007/s12599-015-0391-3

Jarrah, A. I. Z., Yu, G., Krishnamurthy, N., & Rakshit, A. (1993). A Decision Support Framework for Airline Flight Cancellations and Delays. *Transportation Science*. https://doi.org/10.1287/trsc.27.3.266

Kantardzic, M. (2011). Data-Mining Process. In *Data Mining: Concepts, Models, Methods, and Algorithms* (2nd ed.). John Wiley & Sons, Inc. Retrieved from https://books.google.pt/books?hl=pt-PT&lr=&id=ZZ7l6v0CvRMC&oi=fnd&pg=PA1&dq=data+mining+purpose&ots=pNConroIEi&sig=veVVrZamROYP11_4OU98C2Y5XVM&redir_esc=y#v=snippet&q=data mining&f=false

Khanmohammadi, S., Tutun, S., & Kucuk, Y. (2016). A New Multilevel Input Layer Artificial Neural Network for Predicting Flight Delays at JFK Airport. *Procedia Computer Science*, *95*, 237–244. https://doi.org/10.1016/j.procs.2016.09.321

Kim, M. S. (2016). Analysis of short-term forecasting for flight arrival time. *Journal of Air Transport Management*, *52*, 35–41. https://doi.org/10.1016/j.jairtraman.2015.12.002

Kim, Y. J., Choi, S., Briceno, S., & Mavris, D. (2016). A Deep Learning Approach to Flight Delay Prediction. *AIAA/IEEE Digital Avionics Systems Conference - Proceedings*, *December*, 1–6. https://doi.org/10.1109/DASC.2016.7778092

Klein, A., Craun, C., & Lee, R. S. (2010). Airport delay prediction using Weather-Impacted Traffic Index (WITI) model. In *AIAA/IEEE Digital Avionics Systems Conference - Proceedings* (pp. 1–13). https://doi.org/10.1109/DASC.2010.5655493

Knowdelay. (2018). KNOWDELAY. Retrieved January 30, 2018, from http://www.knowdelay.com/how-it-works.html

Koblar, V. (2012). *OPTIMIZING PARAMETERS OF MACHINE LEARNING ALGORITHMS*. Ljubljana.

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference on Articial Intelligence (IJCAI)*, *5*, 1–7. https://doi.org/10.1067/mod.2000.109031

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(12), 273–324.

Kononenko, I., & Kukar, M. (2007a). Measures for Performance Evaluation. In *Machine Learning and Data Mining: Introduction to Principles and Algorithms* (pp. 68–81). Horwood Publishing Limited.

Kononenko, I., & Kukar, M. (2007b). The Name of the Game. In *Machine Learning and Data Mining: Introduction to Principles and Algorithms* (pp. 2–4). Horwood Publishing Limited. Retrieved from https://books.google.pt/books?id=NUikAgAAQBAJ&pg=PA77&hl=pt-PT&source=gbs_selected_pages&cad=3#v=snippet&q=kdd&f=false

Krozel, J., Capozzi, B., Andre, A. D., & Smith, P. (2003). The future National Airspace System: Design requirements imposed by weather constraints. In *AIAA Guidance, Navigation, and Control Conference* (pp. 1–14).

Kumar, G. R., Kongara, V. S., & Ramachandra, D. G. . (2013). An Efficient Ensemble Based Classification Techniques for Medical Diagnosis. *International Journal of Latest Technology in Engineering, Management & Applied Science*, *II*(VIII), 5–9.

Lachheta, P., & Bawa, S. (2016). Combining Synthetic Minority Oversampling Technique and Subset Feature Selection Technique For Class Imbalance Problem. *Proceedings of the International Conference on Advances in Information Communication Technology & Computing - AICTC '16*, (October), 1–6. https://doi.org/10.1145/2979779.2979804

Larose, D. T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining*. John Wiley & Sons, Inc. Retrieved from https://books.google.pt/books?id=JbPMdPWQIOwC&pg=PA28&lpg=PA28&dq=Data+preparation+is+60%25+of+effort+for+data+mining+process+(Pyle)&source=bl&ots=jKmcc9mx5Z&sig=cN-SJ9O5vaX3zGi-amSvguxQmwQ&hl=pt-PT&sa=X&ved=0ahUKEwj3msG29I7XAhWGshQKHa-TBtMQ6AEIQzAE#v=one

Lavalle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big Data, Analytics and the Path From Insights to Value. *MIT Sloan Management Review*, *52*(2), 21–32. https://doi.org/10.0000/PMID57750728

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22. Retrieved from http://www.cs.colorado.edu/department/publications/reports/docs/CU-CS-954-03.pdf%5Cnpapers2://publication/uuid/F28DC17B-D961-4E66-9DEF-B940467A5068%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3223189&tool=pmcentrez&rendertype=abstract%5Cn

Lippmann, R. P. (1987). An Introduction to Computing: with Neural Nets. *IEEE ASSP Magazine*, (April). https://doi.org/10.1109/MASSP.1987.1165576

Liu, X.-Y., Wu, J., & Zhou, Z.-H. (2009). Exploratory Undersampling for Class Imbalance Learning. *IEEE Transactions on Systems, Man and Cybernetics*, *39*(2), 539–550. https://doi.org/10.1109/TSMCB.2008.2007853

Mariscal, G., Marbán, Ó., & Fernández, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *Knowledge Engineering Review*, *25*(2), 137–166. https://doi.org/10.1017/S0269888910000032

Mazzeo, M. J. (2003). Competition and service quality in the U.S. airline industry. *Review of Industrial Organization*, *22*(4), 275–296. https://doi.org/10.1023/A:1025565122721

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill Science/Engineering/Math.

Moisen, G. G. (2008). Classification and Regression Trees. *Encyclopedia of Ecology*, 582–588. https://doi.org/http://www.fs.fed.us/rm/pubs_other/rmrs_2008_moisen_g001.pdf

Mueller, E., & Chatterji, G. (2002). Analysis of aircraft arrival and departure delay characteristics. *AIAA Aircraft Technology, Integration and Operations*. https://doi.org/10.2514/6.2002-5866

Muenchen, R. A. (2017). r4stats.com. Retrieved November 14, 2017, from http://r4stats.com/articles/popularity/

Nadolski, V. L. (1998). *Automated Surface Observing System (ASOS) User's Guide*. Retrieved from http://www.nws.noaa.gov/asos/pdfs/aum-toc.pdf

Office of Aviation Enforcement and Proceedings. (2016). *Air Travel Consumer Reports*. Retrieved from https://www.transportation.gov/airconsumer/air-travel-consumer-reports

Palit, A. K., & Popovic, D. (2005). Neural Networks Approach. In *Computational Intelligence in Time Series Forecasting: Theory and Engineering Applications (Advances in Industrial Control)* (1st ed., p. 372). Springer-Verlag London. https://doi.org/0.1007/1-84628-184-9

Pennsylvania State University. (2017). PennState Eberly College of Science Applied Data Mining and Statistical Learning. Retrieved October 24, 2017, from https://onlinecourses.science.psu.edu/stat857/node/161

Pinkus, A. (1999). Approximation theory of the MLP model in neural networks. *Acta Numerica*, *8*, 143. https://doi.org/10.1017/S0962492900002919

PlanetSpotters.net. (2017). Airline Fleets. Retrieved August 12, 2017, from https://www.planespotters.net/airlines

Pyle, D. (1999). *Data Preparation for Data Mining*. *Order A Journal On The Theory Of Ordered Sets And Its Applications* (Vol. 17). Morgan Kaufmann Publishers, Inc. https://doi.org/10.1080/713827180

Pyrgiotis, N., Malone, K. M., & Odoni, A. (2013). Modelling delay propagation within an airport network. *Transportation Research Part C: Emerging Technologies*, *27*, 60–75. https://doi.org/10.1016/j.trc.2011.05.017

Qianya, L., Lei, W., Rong, F., Bin, W., & Xinhong, H. (2015). An Analysis Method for Flight Delays based on Bayesian Network. *Proceedings of the 2015 27th Chinese Control and Decision Conference, CCDC 2015*, 2561–2565.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Mateo, California: Morgan Kaufmann Publishers, Inc. https://doi.org/10.1007/BF00993309

Rebollo, J. J., & Balakrishnan, H. (2014). Characterization and prediction of air traffic delays. *Transportation Research Part C: Emerging Technologies*, *44*, 231–241. https://doi.org/10.1016/j.trc.2014.04.007

Rong, F., Qianya, L., Bo, H., Jing, Z., & Dongdong, Y. (2015). The Prediction of Flight Delays based the Analysis of Random Flight Points. *34th Chinese Control Conference (CCC)*, 3992–3997. https://doi.org/10.1109/ChiCC.2015.7260255

Saeys, Y., Inza, I., & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*(19), 2507–2517. https://doi.org/10.1093/bioinformatics/btm344

SAS. (2017). Machine Learning: O que é e por que é importante? Retrieved July 17, 2017, from https://www.sas.com/pt_br/insights/analytics/machine-learning.html

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press. Retrieved from https://books.google.pt/books?hl=pt-PT&lr=&id=Hf6QAwAAQBAJ&oi=fnd&pg=PR15&dq=Understanding+machine+learning:+from+theory+to+algorithms.+2014&ots=2HqkSihNM7&sig=Y76uKAFNYuWYLT4RmsTxr20SJck&redir_esc=y#v=onepage&q=neural networks&f=false

Smallen, D. (2016). *2015 U. S. Based Airline Traffic Data*. Retrieved from http://www.rita.dot.gov/bts/sites/rita.dot.gov.bts/files/bts18_16.pdf

Smartertravel. (2009). SMARTERTRAVEL. Retrieved January 30, 2018, from https://www.smartertravel.com/2009/08/19/new-tool-predicts-flight-delays/

Smith, D. a, & Sherry, L. (2008). Decision Support Tool for Predicting Aircraft Arrival Rates , Ground Delay Programs , and Airport Delays from Weather Forecasts. *Systems Research*. https://doi.org/10.1109/ICNSURV.2008.4559186

Soley-bori, M. (2013). *Dealing with missing data: Key assumptions and methods for applied analysis*. *Technical Report No. 4*. Retrieved from https://www.bu.edu/sph/files/2014/05/Marina-tech-report.pdf

Sternberg, A., Soares, J., Carvalho, D., & Ogasawara, E. (2017). A Review on Flight Delay Prediction, 1–15. Retrieved from http://arxiv.org/abs/1703.06118

Takacs, G. (2014). Predicting flight arrival times with a multistage model. *2014 IEEE International Conference on Big Data, IEEE Big Data 2014*, *2*, 78–84. https://doi.org/10.1109/BigData.2014.7004435

Tang, J., Alelyani, S., & Liu, H. (2014). Feature Selection for Classification: A Review. In C. Aggarwal (Ed.), *Data Classification: Algorithms and Applications* (pp. 37–64). CRC Press. https://doi.org/10.1.1.409.5195

Teorey, T. J., Yang, D., & Fry, J. P. (1986). A Logical Design Methodology for Relational Databases Using the Extended Entity-Relationship Model. *ACM Computing Surveys*, *18*(2), 197–222. https://doi.org/10.1145/7474.7475

Tourism Review. (2008). Tourism Review. Retrieved January 30, 2018, from https://www.tourism-review.com/delaycast-predicting-flight-delays-news856

TravelMath. (2017a). TravelMath. Retrieved November 17, 2017, from https://www.travelmath.com/time-change/

TravelMath. (2017b). TravelMath. Retrieved November 17, 2017, from https://www.travelmath.com/flying-time/

Tryfona, N., Busborg, F., & Christiansen, J. G. B. (1999). starER: A Conceptual Model for Data Warehouse Design. *ACM 2nd International Workshop Data Warehousing and OLAP*, 3–8. https://doi.org/10.1145/319757.319776

Tu, Y., Ball, M. O., & Jank, W. S. (2008a). Estimating Flight Departure Delay Distributions—A Statistical Approach With Long-Term Trend and Short-Term Pattern. *Journal of the American Statistical Association*, *103*(481), 112–125. https://doi.org/10.1198/016214507000000257

Tu, Y., Ball, M. O., & Jank, W. S. (2008b). Estimating Flight Departure Delay Distributions—A Statistical Approach With Long-Term Trend and Short-Term Pattern. *Journal of the American Statistical Association*, *103*(481), 112–125. https://doi.org/10.1198/016214507000000257

Tukey, J. W. (1961). *The Future of Data Analysis*. *Annals of the Institute of Statistical Mathematics*. https://doi.org/10.1214/aoms/1177728422

U.S. Department of Transportation. (2017). 2015 Flight Delays and Cancellations: Which airline should you fly on to avoid significant delays? Retrieved August 12, 2017, from

https://www.kaggle.com/usdot/flight-delays

U.S. Department of Transportation, Federal Aviation Administration, & Corporation, M. (2004). *Airport Capacity Benchmark Report 2004*. Washington, D.C. Retrieved from ftp://ftp.agl.faa.gov/ORD DEIS/Referance Documents/Appendix A/App A - Ref Doc 10.pdf

U.S. Office of Personnel Management. (2017a). OPM.GOV. Retrieved November 17, 2017, from https://www.opm.gov/about-us/our-mission-role-history/what-we-do/

U.S. Office of Personnel Management. (2017b). OPM.GOV. Retrieved September 8, 2017, from https://www.opm.gov/policy-data-oversight/snow-dismissal-procedures/federal-holidays/

University of Waikato. (2018). Weka Wikispace. Retrieved January 9, 2018, from https://weka.wikispaces.com/

Wang, C. W. (2006). New Ensemble Machine Learning Method for Classification and Prediction on Gene Expression Data. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society.*, *1*(Aug-Sept), 3478–81. https://doi.org/10.1109/IEMBS.2006.259893

Wang, X. W. X. (2009). Intelligent Quality Management Using Knowledge Discovery in Databases. *2009 International Conference on Computational Intelligence and Software Engineering*, 1–4. https://doi.org/10.1109/CISE.2009.5364999

Weiss, G. M. (2004). Mining with Rarity: A Unifying Framework. *SIGKDD Explorations*, *6*(1), 7–19. https://doi.org/10.1145/1007730.1007734

Witten, I. H., Frank, E., & Hall, M. a. (2011a). *Data Mining: Practical Machine Learning Tools and Techniques* (2nd ed.). Morgan Kaufmanne. https://doi.org/0120884070, 9780120884070

Witten, I. H., Frank, E., & Hall, M. A. (2011b). *Data mining: Practical Machine Learning Tools and Techniques, Third Edition*. *Morgan Kaufmann series in data management systems* (3rd ed.). Elsevier Inc. https://doi.org/10.1002/1521-3773(20010316)40:6<9823::AID-ANIE9823>3.3.CO;2-C

Wu, S.-H., & DataLab. (2016). Cross Validation & Ensembling. Retrieved October 25, 2017, from http://www.cs.nthu.edu.tw/~shwu/courses/ml/labs/08_CV_Ensembling/08_CV_Ensembling.html

Xu, N., Donohue, G., Laskey, K. B., & Chen, C. (2005). Estimation of Delay Propagation in the National Aviation System Using Bayesian Networks. *6th USA/Europe Air Traffic Management Research and Development Seminar*, 478–489. Retrieved from https://pdfs.semanticscholar.org/9c5a/7c726315387acb5ff5cf8e849524046cd207.pdf

Yablonsky, G., Steckel, R., Constales, D., Farnan, J., Lercel, D., & Patankar, M. (2014). Flight delay performance at Hartsfield-Jackson Atlanta International Airport. *Journal of Airline and Airport Management*, *4*(1), 78–95. https://doi.org/10.3926/jairm.22

Yao, R., Jiandong, W., & Tao, X. (2010). A flight delay prediction model with consideration of cross-flight plan awaiting resources. In *Proceedings - 2nd IEEE International Conference on Advanced Computer Control, ICACC* (Vol. 4, pp. 1–5). https://doi.org/10.1109/ICACC.2010.5487088

Zhang, G., Patuwo, B. E., & Hu, M. Y. (1997). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, *14*, 35–62. https://doi.org/10.1016/S0169-2070(97)00044-7

Zhang, S., Yang, Q., & Zhang, C. (2002). Proceedings of the First International Workshop on Data Cleaning and Preprocessing (p. iii). Maebashi, Japan. Retrieved from https://www.researchgate.net/profile/Stefano_Cerri/publication/232866130_GAsRULE_for_Knowledge_Discovery/links/09e4150c6f1905948f000000/GAsRULE-for-Knowledge-Discovery.pdf

Zonglei, L., Jiandong, W., & Guansheng, Z. (2008). A New Method to Alarm Large Scale of Flights Delay Based on Machine Learning. In *Proceedings of the International Symposium on Knowledge Acquisition and Modeling (KAM 08)* (pp. 589–592). Wuhan, China: IEEE CS Press. https://doi.org/10.1109/KAM.2008.18

Zonglei, L., Jiandong, W., & Tao, X. (2009). A New Method for Flight Delays Forecast Based on the Recommendation System. *ISECS International Colloquium on Computing, Communication, Control and Management*, *1*, 46–49. https://doi.org/10.1109/CCCM.2009.5268153

Zupan, B., & Demsar, J. (2008). Open-Source Tools for Data Mining. *Clinics in Laboratory Medicine*, *28*(1), 37–54. https://doi.org/10.1016/j.cll.2007.10.002

8. ANNEXES

ANNEX 1: PROCEEDINGS FOR DATA VALIDATION BY TYPE OF SOURCE DATA INFORMATION

For the integration, each source information had to be treated to provide homogeneity as is explained in this annex.

A. Flight Information

The data directly extracted from the BTS with information about the flights was not in the right format. Some of the operations made were: (1) representing the minutes as a unit, instead of five minutes the values were represented as 5.00 minutes, for that reason all variables with minute values were transformed from text to number for a better understanding; (2) to be able to understand the real time, because the variables that represented hours were in total of minutes (90) and not in hour format (01:30), the second format was applied to all varibales to which this decision could be applied. This decision also helped to make possible the calculations of hours for the construction of further variables and for consulting the database.

Due to a large amount of data extracted per month, approximately more than twenty thousand flights, it would not be possible to make these changes manually in each extracted excel file. Therefore, the solution of these problems was possible through VBA code in the monthly files.

B. Airplane Information

To obtain information about the airplane of each flight, the FAA website was used (Federal Aviation Administration, 2016a). As an airplane can be used on more than one flight in a month, first all the aircraft used in the eight months in question were identified, representing a total of 3279 different aircraft.

Then, it was necessary to confirm the validity of the unique number of the aircraft, the N-Number. Following specific rules of format, stipulated by the FAA, an N-Number can contain:

- One to five numbers (e.g., N12345);
- One to four numbers followed by one letter (e.g., N1234Z);
- One to three number followed by two letters (e.g., N123AZ).

On the other hand, it cannot contain a zero as the first number (e.g., N01234) or the letters "i" or "o" (e.g., N1234I or N123AO) because they can be mistaken for numbers one and zero (Federal Aviation Administration, 2015).

After confirmation, only two thousand and nine hundred sixty-five (2965) airplanes had a correct N-Number. The remaining three thousand and fourteen (314) aircraft wrongly presented another type of indicator number, the so-called Fleet Number - corresponding to the number of the airplane within the airline -, often supplied to the BTS as the N-Number. Therefore, to obtain the N-Number, registered by the NAA referring to the Fleet Number present in the flight data, it was necessary to resort to the database of these numbers of each airline through the use of the *PlanetSpotters* website, a civil database with airplanes information (PlanetSpotters.net, 2017). After accessing the database with information about each airline's Fleet Number, we concluded that from the three thousand and fourteen (314) records to be validated was not possible to identify the N-Number of

seven (7), thus assuming the number presented in the flight data as invalid and consequently, flight showing these N-Numbers were not included in the final dataset.

Finally, it was necessary to manually search and fill information of each airplane by consulting the N-number (Federal Aviation Administration, 2016c) and model airplane (Federal Aviation Administration, 2016b) on the FAA website.

For authenticate the fulfillment of the FAA registration rules of the N-Number, and due to the large amount of data, we used Excel functions.

C. Weather Information

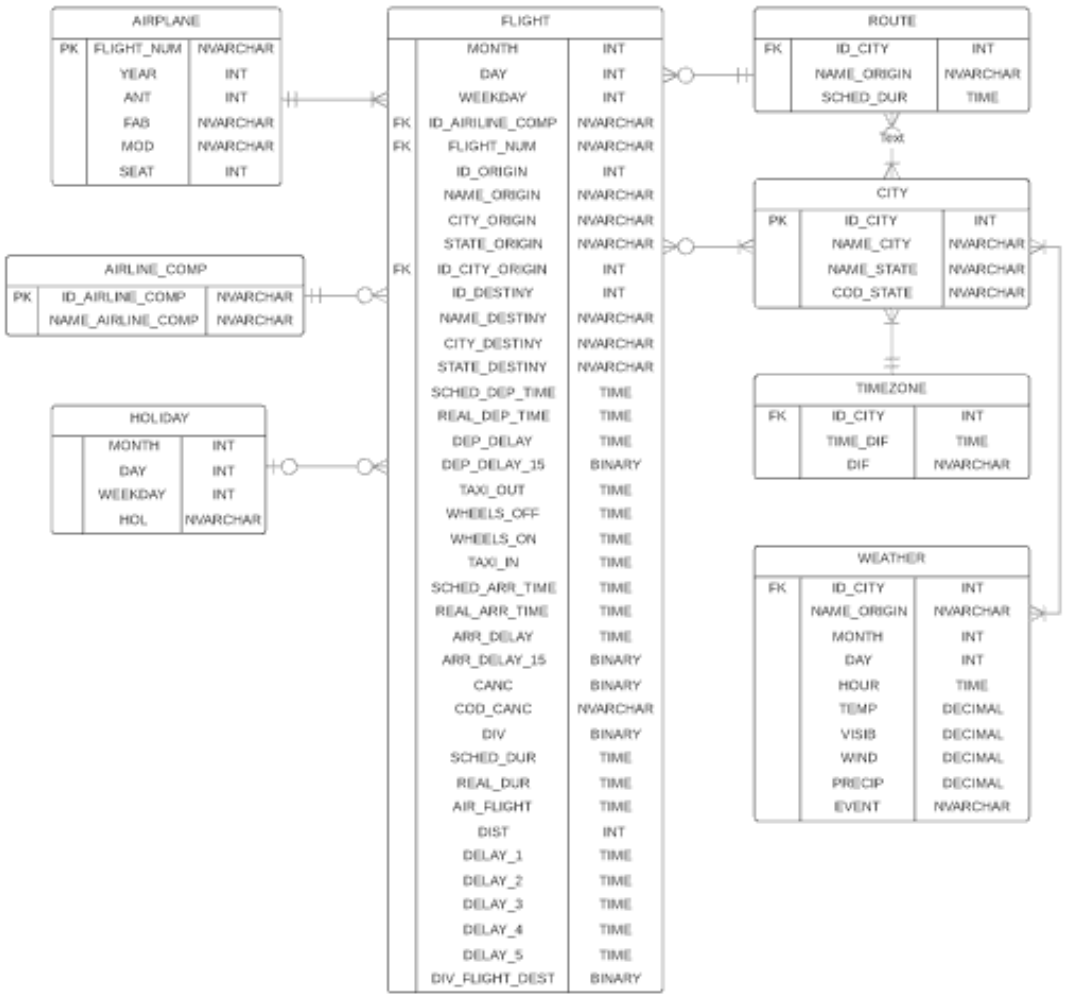
The acquisition of weather information was the one which causes more difficult regarding availability because it was hard to find a repository with historical information of all the airport's locations. In that sense, was found the IEM, where the data was stored in the archive and was necessary to download them. The download was done for the year of 2015 and for all the fifty states, because the website allowed to download the information of multiple airports in the same state at once.

It was obtained the hourly observations about temperature (in Celsius degrees), precipitation (in millimeters per hour), wind speed (in miles per hour), visibility (in miles) and weather phenomena event at the time of observation (according to METAR weather phenomena that are reported in terms of types and characteristics, and qualified with respect to intensity or proximity to the aerodrome (International Civil Aviation Organization, 2011)).

During this phase some airports sensors did not contain historical data of certain hours of the day and a great part of the present weather information was represented with a value “M”, being mentioned by the website as an observation that was reported as missing or that were set to missing after quality control check, or a value that was never reported by the ASOS sensor. This kind of problems may arise, for example, for some flight in a certain hour (that was not available) a missing value in variables of weather.

After extraction of the historical hourly data for each day of the year of 2015 for all one hundred sixty-nine (169) airports it was necessary the treatment of each excel of each state resorting to excel functions. Duplicate records elimination was necessary as well as the correction of information of present weather codes assumed by the excel as functions and not text information.

ANNEX 2: ENTITY-RELATIONSHIP PHYSICAL MODEL WITH *CROW'S FOOT* NOTATION



ANNEX 3: DISTINCT AIRPORTS AND CITIES WITH RESPECT TO TIME-ZONE

Airport Name	*City and State*	*Time-Zone in relation to Atlanta*
ATL	Atlanta, Georgia	-------
DFW	Dallas/Fort Worth, Texas	-01:00
MIA	Miami, Florida	00:00
SEA	Seattle, Washington	-03:00
GSO	Greensboro/High Point, North Carolina	00:00
LAX	Los Angeles, California	-03:00
MDW	Chicago, Illinois	-01:00
ORD		
JFK	New York, New York	00:00
LGA		
SJU	San Juan, Puerto Rico	00:00

SAT	San Antonio, Texas	-01:00
RSW	Fort Myers, Florida	00:00
DAY	Dayton, Ohio	00:00
AVL	Asheville, North Carolina	00:00
SDF	Louisville, Kentucky	00:00
EWR	Newark, New Jersey	00:00
STT	Charlotte Amalie, U.S. Virgin Islands	00:00
OMA	Omaha, Nebraska	-01:00
JAX	Jacksonville, Florida	00:00
PIT	Pittsburgh, Pennsylvania	00:00
CMH	Columbus, Ohio	00:00
MEM	Memphis, Tennessee	-01:00
PBI	West Palm Beach/Palm Beach, Florida	00:00
MKE	Milwaukee, Wisconsin	-01:00
MSY	New Orleans, Louisiana	-01:00
MSP	Minneapolis, Minnesota	-01:00
MCO	Orlando, Florida	00:00
TPA	Tampa, Florida	00:00
RDU	Raleigh/Durham, North Carolina	00:00
SNA	Santa Ana, California	-03:00
PDX	Portland, Oregon	-03:00
PHL	Philadelphia, Pennsylvania	00:00
HOU	Houston, Texas	-01:00
IAH		
CLT	Charlotte, North Carolina	00:00
FLL	Fort Lauderdale, Florida	00:00
MOB	Mobile, Alabama	-01:00
HNL	Honolulu, Hawaii	-05:00
DCA	Washington DC, Virginia	00:00
IAD		
TLH	Tallahassee, Florida	00:00
JAN	Jackson/Vicksburg, Mississippi	-01:00
GRR	Grand Rapids, Michigan	00:00
DTW	Detroit, Michigan	00:00
EYW	Key West, Florida	00:00
BDL	Hartford, Connecticut	00:00
MCI	Kansas City, Missouri	-01:00
BOS	Boston, Massachusetts	00:00
LAS	Las Vegas, Nevada	-03:00
PNS	Pensacola, Florida	-01:00
STL	St. Louis, Missouri	-01:00

DAB	Daytona Beach, Florida	00:00
IND	Indianapolis, Indiana	00:00
ATW	Appleton, Wisconsin	-01:00
PVD	Providence, Rhode Island	00:00
AUS	Austin, Texas	-01:00
BHM	Birmingham, Alabama	-01:00
BUF	Buffalo, New York	00:00
SAV	Savannah, Georgia	00:00
EGE	Eagle, Colorado	-02:00
RIC	Richmond, Virginia	00:00
ORF	Norfolk, Virginia	00:00
DEN	Denver, Colorado	-02:00
PHX	Phoenix, Arizona	-02:00
SFO	San Francisco, California	-03:00
SYR	Syracuse, New York	00:00
TUS	Tucson, Arizona	-02:00
GPT	Gulfport/Biloxi, Mississippi	-01:00
SLC	Salt Lake City, Utah	-02:00
BNA	Nashville, Tennessee	-01:00
GSP	Greer, South Carolina	00:00
JAC	Jackson, Wyoming	-02:00
ROC	Rochester, New York	00:00
SAN	San Diego, California	-03:00
TRI	Bristol/Johnson City/Kingsport, Tennessee	00:00
AVP	Scranton/Wilkes-Barre, Pennsylvania	00:00
BWI	Baltimore, Maryland	00:00
CLE	Cleveland, Ohio	00:00
LIT	Little Rock, Arkansas	-01:00
CVG	Cincinnati, Kentucky	00:00
MHT	Manchester, New Hampshire	00:00
ECP	Panama City, Florida	-01:00
ABQ	Albuquerque, New Mexico	-02:00
MLB	Melbourne, Florida	00:00
CHS	Charleston, South Carolina	00:00
ALB	Albany, New York	00:00
FNT	Flint, Michigan	00:00
VPS	Valparaiso, Florida	-01:00
TYS	Knoxville, Tennessee	00:00
COS	Colorado Springs, Colorado	-02:00
ELP	El Paso, Texas	-02:00
SRQ	Sarasota/Bradenton, Florida	00:00

CAK	Akron, Ohio	00:00
MDT	Harrisburg, Pennsylvania	00:00
OKC	Oklahoma City, Oklahoma	-01:00
HSV	Huntsville, Alabama	-01:00
SMF	Sacramento, California	-03:00
HDN	Hayden, Colorado	-02:00
MTJ	Montrose/Delta, Colorado	-02:00
ICT	Wichita, Kansas	-01:00
OAK	Oakland, California	-03:00
CHO	Charlottesville, Virginia	00:00
LEX	Lexington, Kentucky	00:00
MSO	Missoula, Montana	-02:00
PWM	Portland, Maine	00:00
BZN	Bozeman, Montana	02:00
ROA	Roanoke, Virginia	00:00
FAY	Fayetteville, North Carolina	00:00
SGF	Springfield, Missouri	01:00
MYR	Myrtle Beach, South Carolina	00:00
ONT	Ontario, California	-03:00
DAL	Dallas, Texas	-01:00
LFT	Lafayette, Louisiana	-01:00
SJC	San Jose, California	-03:00
CAE	Columbia, South Carolina	00:00
MSN	Madison, Wisconsin	-01:00
TUL	Tulsa, Oklahoma	-01:00
AGS	Augusta, Georgia	00:00
STX	Christiansted, U.S. Virgin Islands	00:00
GNV	Gainesville, Florida	00:00
XNA	Fayetteville, Arkansas	-01:00
ABE	Allentown/Bethlehem/Easton, Pennsylvania	00:00
CRW	Charleston/Dunbar, West Virginia	00:00
CHA	Chattanooga, Tennessee	00:00
DSM	Des Moines, Iowa	-01:00
ILM	Wilmington, North Carolina	00:00
EVV	Evansville, Indiana	-01:00
SHV	Shreveport, Louisiana	-01:00
BTR	Baton Rouge, Louisiana	-01:00
SBN	South Bend, Indiana	00:00
GRB	Green Bay, Wisconsin	-01:00
CID	Cedar Rapids/Iowa City, Iowa	-01:00
ASE	Aspen, Colorado	-02:00

MGM	Montgomery, Alabama	-01:00
RST	Rochester, Minnesota	-01:00
FWA	Fort Wayne, Indiana	00:00
HPN	White Plains, New York	00:00
BTV	Burlington, Vermont	00:00
GTR	Columbus, Mississippi	-01:00
FAR	Fargo, North Dakota	-01:00
DHN	Dothan, Alabama	-01:00
PHF	Newport News/Williamsburg, Virginia	00:00
VLD	Valdosta, Georgia	00:00
FSM	Fort Smith, Arkansas	-01:00
PIA	Peoria, Illinois	-01:00
BMI	Bloomington/Normal, Illinois	-01:00
MLU	Monroe, Louisiana	-01:00
BQK	Brunswick, Georgia	00:00
CSG	Columbus, Georgia	00:00
OAJ	Jacksonville/Camp Lejeune, North Carolina	00:00
AEX	Alexandria, Louisiana	-01:00
LNK	Lincoln, Nebraska	-01:00
ELM	Elmira/Corning, New York	00:00
LAN	Lansing, Michigan	00:00
MLI	Moline, Illinois	-01:00
EWN	New Bern/Morehead/Beaufort, North Carolina	00:00
GRK	Killeen, Texas	-01:00
ABY	Albany, Georgia	00:00
RAP	Rapid City, South Dakota	-02:00
MBS	Saginaw/Bay City/Midland, Michigan	00:00
AZO	Kalamazoo, Michigan	00:00
SCE	State College, Pennsylvania	00:00
FSD	Sioux Falls, South Dakota	-01:00
TTN	Trenton, New Jersey	00:00
ACY	Atlantic City, New Jersey	00:00
ANC	Anchorage, Alaska	-04:00
FCA	Kalispell, Montana	-02:00
TVC	Traverse City, Michigan	00:00

ANNEX 4: AIRLINE COMPANIES

Airline Company ID	*Airline Company Name*
AA	American Airlines Inc.
AS	Alaska Airlines Inc.

DL	Delta Air Lines Inc.
NK	Spirit Airlines
OO	SkyWest Airlines
EV	ExpressJet Airlines Inc
UA	United Airlines Inc.
MQ	Envoy Air Inc.
US	US Airways
WN	Southwest Airlines Co.
F9	Frontier Airlines Inc.

ANNEX 5: FEDERAL HOLIDAYS

Celebrated date in 2015 [8]	*Holiday*	*Description*
January 1st	New Year's Day	-
January 19th	Birthday of Martin Luther King, Jr.	National holiday in honor of Martin Luther King Jr, made official in 1983. It is celebrated on the third Monday of the month of January
February 16th	Washington's Birthday/ *Presidents' Day*	National holiday in honor of George Washington, first president of the USA, as well as to the current president to reside. It is celebrated on the third Monday of February
May 25th	*Memorial Day*	National holiday in honor of the American military who died in combat. It is celebrated on the last Monday of May
July 3th	Independence Day	National holiday marking the declaration of independence. It is celebrated on July 4 (Saturday)
September 7th	Labor Day	National holiday celebrating the social and economic contribution of workers to the country. It is celebrated on the first Monday of September
October 12th	Columbus Day	National holiday celebrating the arrival of Christopher Columbus to America. It is celebrated on October 12
November 11th	Veterans Day	Holiday honoring the veteran military. It is celebrated on November 11
November 26th	Thanksgiving Day	National holiday celebrated as a way of thanking the events of the year in question. It is celebrated on the fourth Thursday of November
December 25th	Christmas Day	-

[8] In accordance to U.S. Office of Personnel Management (2017b), if a holiday occurs on a Saturday or Sunday, automatically, the same holiday will be celebrated in Friday or Monday, respectively.

ANNEX 6: ACQUISITION OF THE 3 PHASES OF THE WEATHER VARIABLES

For the creation of the weather variables it was needed to think in the three phases that could influence a flight to decollate or land.

For this reason, the explanation in detail on how we achieved these 3 phases values for each weather variable (Temperature, Precipitation, Wind, Visibility and Event) is explained below.

- **At the origin, at the scheduled time of departure**, attainable through the querying of the hour of the scheduled departure time of the flight table on the weather table;
- **At the destination, at the scheduled time of departure in origin, with the time zone of the destination**, achievable through the creation of a variable with the time zone, between the origin and destination, applied to the scheduled departure time. Obtained by querying the weather table and seeing, at the equivalent schedule departure time from the origin, on the destination (departure time with time zone applied), the information needed. Sometimes the time zone, when applied, could change the day at the destination. For that reason, as a way to make the calculation of this variable possible, it was also necessary to create variables for the day and the month that could be altered by the time zone only to consult the right day and month in the destination. For example, to know the wind speed at the destination, Hartsfield-Jackson International Airport (ATL), on a flight from Dallas/Fort Worth International Airport (DFW) with a scheduled time of departure at 23:30 on January 10, we have to check the wind speed records in the city of ATL airport on January 11 at 00:30, due to the time zone that is one hour ahead from DFW to ATL. This type of situation is an option not yet considered in previous studies.
- **At the destination, at the scheduled time of arrival**, available through the querying of the hour of the scheduled arrival time of the flight table on the weather table.

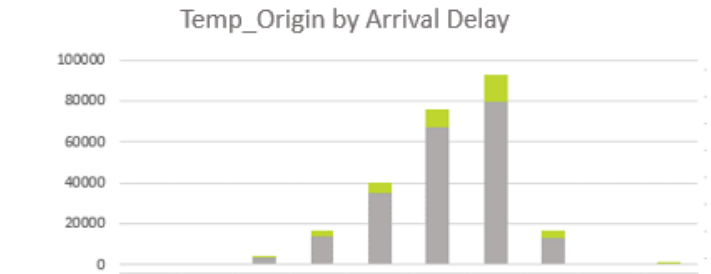
ANNEX 7: INFLUENCE OF ARRIVAL DELAY ON TEMPERATURE VARIABLES

In each graph is contained two types of information:

- the frequency of each class by the dependent variable on the vertical axis
- the relevance in percentage of each class of the dependent variable by each class of the independent variable in analysis on the horizontal axis

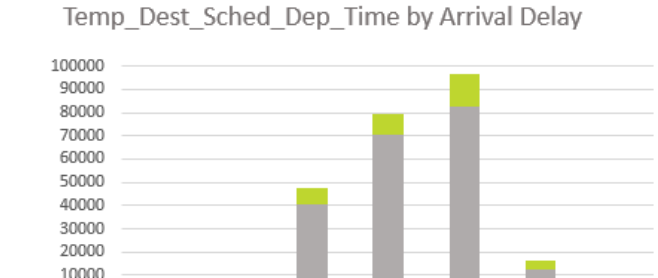
The next figures show the three variables about the temperature where is illustrated:

A. Influence of Arrival Delay on Temperature at the origin at the schedule departure time



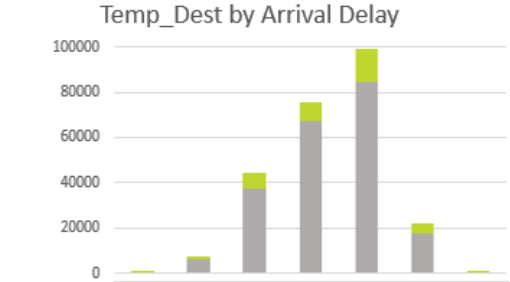
	< -30°C	[-30°C; -20°C[	[-20°C; -10°C[	[-10°C; 0°C[	[0°C; 10°C[	[10°C; 20°C[	[20°C; 30°C[	[30°C; 40°C[	[40°C; 50°C[	Missing
Delayed	100,0%	19,8%	19,8%	18,3%	12,3%	11,4%	14,4%	23,9%	21,9%	13,4%
Not Delayed	0,0%	80,2%	80,2%	81,7%	87,7%	88,6%	85,6%	76,1%	78,1%	86,6%

B. Influence of Arrival Delay on Temperature at the destination at the schedule departure time in the origin



	< -10°C	[-10°C; 0°C[	[0°C; 10°C[	[10°C; 20°C[	[20°C; 30°C[	[30°C; 40°C[	Missing
Delayed	10,7%	12,7%	14,7%	11,3%	15,0%	22,7%	6,3%
Not Delayed	89,3%	87,3%	85,3%	88,7%	85,0%	77,3%	93,7%

C. Influence of Arrival Delay on Temperature at the destination at the schedule arrival time



	< -10°C	[-10°C; 0°C[	[0°C; 10°C[	[10°C; 20°C[	[20°C; 30°C[	[30°C; 40°C[	Missing
Delayed	11,4%	12,5%	15,3%	11,4%	14,8%	19,7%	8,8%
Not Delayed	88,6%	87,5%	84,7%	88,6%	85,2%	80,3%	91,2%

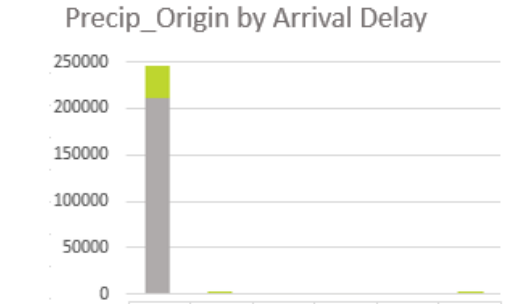
ANNEX 8: INFLUENCE OF ARRIVAL DELAY ON PRECIPITATION VARIABLES

In each graph is contained two types of information:

- the frequency of each class by the dependent variable on the vertical axis
- the relevance in percentage of each class of the dependent variable by each class of the independent variable in analysis on the horizontal axis

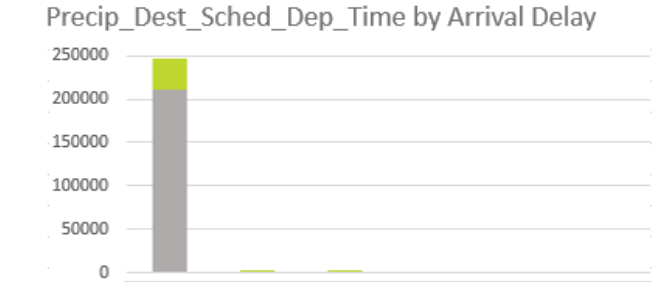
The next figures show the three variables about the precipitation where is illustrated:

A. Influence of Arrival Delay on Precipitation at the origin at the schedule departure time



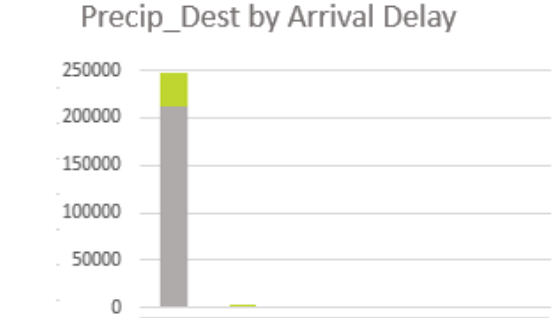
	[0;2[	[2;5[	[5;10[	[10;50[	≥ 50	Missing
Delayed	14,1%	23,2%	27,2%	33,3%	0,0%	12,6%
Not Delayed	85,9%	76,8%	72,8%	66,7%	100,0%	87,4%

B. Influence of Arrival Delay on Precipitation at the destination at the schedule departure time in the origin



	[0;2[	[2;5[	[5;10[	[10;50[	≥ 50	Missing
Delayed	14,1%	24,2%	15,5%	62,9%	0,0%	6,3%
Not Delayed	85,9%	75,8%	84,5%	37,1%	0,0%	93,7%

C. Influence of Arrival Delay on Precipitation at the destination at the schedule arrival time



	[0;2[	[2;5[	[5;10[	[10;50[	≥ 50	Missing
Delayed	14,1%	19,8%	23,6%	62,9%	0,0%	9,0%
Not Delayed	85,9%	80,2%	76,4%	37,1%	0,0%	91,0%

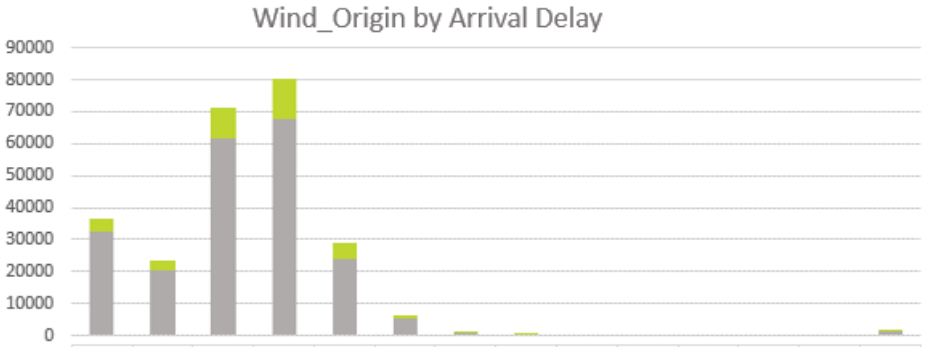
ANNEX 9: INFLUENCE OF ARRIVAL DELAY ON WIND VARIABLES

In each graph is contained two types of information:

- the frequency of each class by the dependent variable on the vertical axis
- the relevance in percentage of each class of the dependent variable by each class of the independent variable in analysis on the horizontal axis

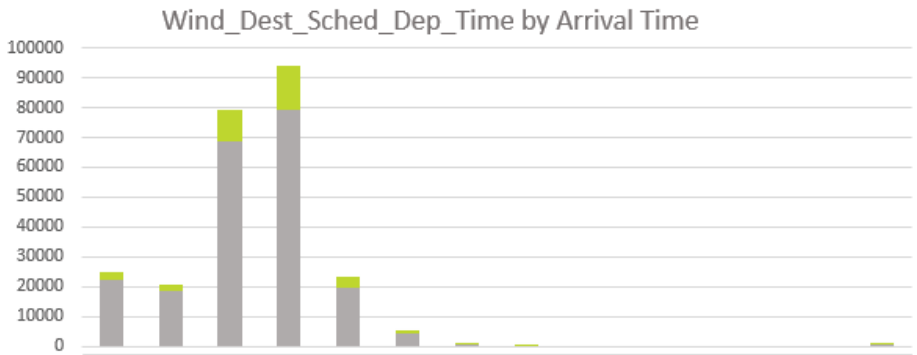
The next figures show the three variables about the wind where is illustrated:

A. Influence of Arrival Delay on Wind at the origin at the schedule departure time



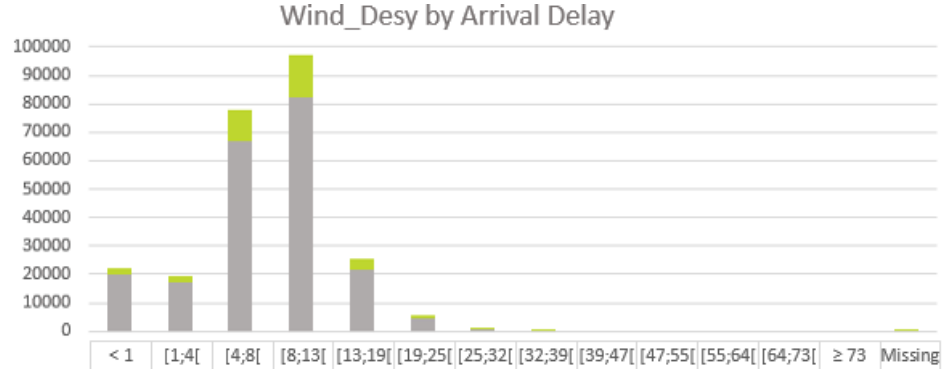
	< 1	[1;4[	[4;8[	[8;13[	[13;19[	[19;25[	[25;32[	[32;39[	[39;47[	[47;55[	[55;64[	[64;73[	≥ 73	Missing
Delayed	10,7%	12,2%	13,3%	15,4%	17,4%	19,9%	20,5%	22,5%	12,5%	0,0%	0,0%	0,0%	0,0%	15,1%
Not Delayed	89,3%	87,8%	86,7%	84,6%	82,6%	80,1%	79,5%	77,5%	87,5%	0,0%	0,0%	0,0%	100,0%	84,9%

B. Influence of Arrival Delay on Wind at the destination at the schedule departure time in the origin



	< 1	[1;4[	[4;8[	[8;13[	[13;19[	[19;25[	[25;32[	[32;39[	[39;47[	[47;55[	[55;64[	[64;73[	≥ 73	Missing
Delayed	10,2%	10,8%	13,7%	16,0%	15,4%	16,7%	16,0%	17,0%	0,0%	0,0%	0,0%	0,0%	0,0%	7,6%
Not Delayed	89,8%	89,2%	86,3%	84,0%	84,6%	83,3%	84,0%	83,0%	0,0%	0,0%	0,0%	0,0%	0,0%	92,4%

C. Influence of Arrival Delay on Wind at the destination at the schedule arrival time



	< 1	[1;4[	[4;8[	[8;13[	[13;19[	[19;25[	[25;32[	[32;39[	[39;47[	[47;55[	[55;64[	[64;73[	≥ 73	Missing
Delayed	9,9%	11,2%	13,9%	15,4%	16,0%	15,2%	17,7%	34,0%	0,0%	0,0%	0,0%	0,0%	0,0%	10,9%
Not Delayed	90,1%	88,8%	86,1%	84,6%	84,0%	84,8%	82,3%	66,0%	0,0%	0,0%	0,0%	0,0%	0,0%	89,1%

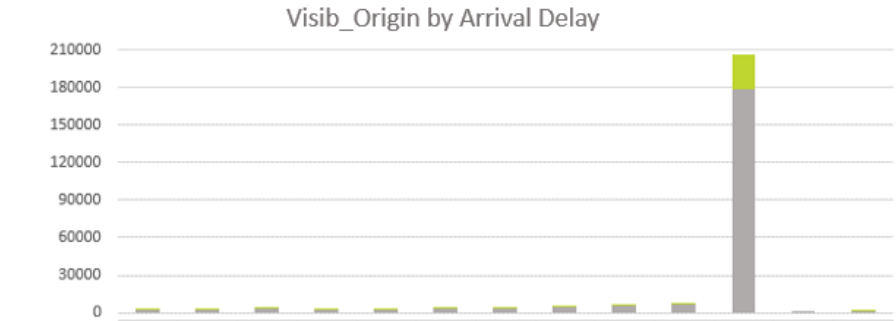
ANNEX 10: INFLUENCE OF ARRIVAL DELAY ON VISIBILITY VARIABLES

In each graph is contained two types of information:

- the frequency of each class by the dependent variable on the vertical axis
- the relevance in percentage of each class of the dependent variable by each class of the independent variable in analysis on the horizontal axis

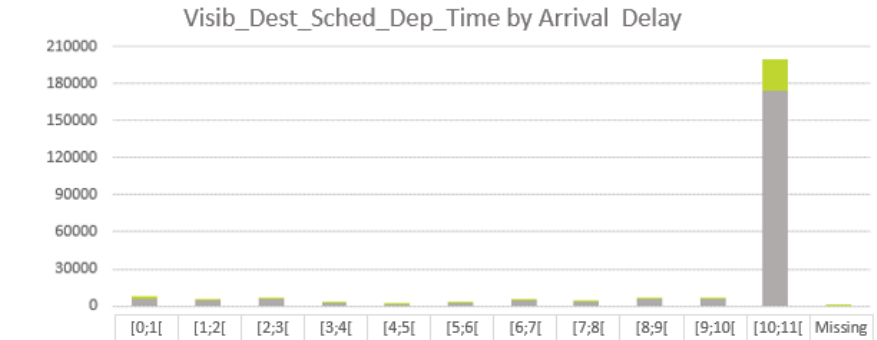
The next figures show the three variables about the visibility where is illustrated:

A. Influence of Arrival Delay on Visibility at the origin at the schedule departure time



	[0;1[	[1;2[	[2;3[	[3;4[	[4;5[	[5;6[	[6;7[	[7;8[	[8;9[	[9;10[	[10;11[	≥11	Missing
Delayed	27,0%	28,3%	24,8%	21,4%	19,2%	18,1%	16,3%	16,2%	15,1%	14,4%	13,3%	0,0%	13,6%
Not Delayed	73,0%	71,7%	75,2%	78,6%	80,8%	81,9%	83,7%	83,8%	84,9%	85,6%	86,7%	100,0%	86,4%

B. Influence of Arrival Delay on Visibility at the destination at the schedule departure time in the origin



	[0;1[	[1;2[	[2;3[	[3;4[	[4;5[	[5;6[	[6;7[	[7;8[	[8;9[	[9;10[	[10;11[	Missing
Delayed	27,3%	16,5%	23,3%	17,9%	24,1%	20,4%	18,6%	15,6%	17,1%	20,5%	12,7%	6,3%
Not Delayed	72,7%	83,5%	76,7%	82,1%	75,9%	79,6%	81,4%	84,4%	82,9%	79,5%	87,3%	93,7%

C. Influence of Arrival Delay on Visibility at the destination at the schedule arrival time

	[0;1[	[1;2[	[2;3[	[3;4[	[4;5[	[5;6[	[6;7[	[7;8[	[8;9[	[9;10[	[10;11[	Missing
Delayed	26,6%	18,7%	22,5%	20,4%	25,7%	23,3%	18,0%	16,2%	19,3%	19,6%	12,6%	9,0%
Not Delayed	73,4%	81,3%	77,5%	79,6%	74,3%	76,7%	82,0%	83,8%	80,7%	80,4%	87,4%	91,0%

ANNEX 11: TEST RESULTS TABLE OF SMOTE TECHNIQUE

Dataset	Sampling Technique	Variable *Dep_Delay + Real_Dep_Time* Presence	Outliers Presence	Attribute Selection		Algorithm	CCI	PR	RE	F	ROC	T to build model	T to test model
FlightData	SMOTE	With	With	CfsSubsetEval	6	J48	79,7689	0,892	0,798	0,824	0,833	10,29	0,15
						RF	79,5948	0,892	0,796	0,823	0,828	132,6	3,04
						MLP	79,7006	0,892	0,797	0,823	0,838	121,76	0,19
				GainRatio	10	J48	79,7716	0,892	0,798	0,824	0,835	26,76	0,5
						RF	79,7716	0,892	0,798	0,824	0,826	172,35	3,63
						MLP	54,1045	0,863	0,541	0,601	0,878	197,16	0,41
					15	J48	79,5975	0,892	0,796	0,823	0,834	31,7	0,21
						RF	79,7649	0,892	0,798	0,824	0,834	217,14	3,58
						MLP	79,7247	0,892	0,797	0,824	0,89	374,49	0,57
					20	J48	72,2549	0,882	0,723	0,763	0,762	45,88	0,32
						RF	79,7609	0,892	0,798	0,824	0,832	236,38	2,65
						MLP	14,3653	0,877	0,144	0,037	0,86	464,76	0,49
			Without	CfsSubsetEval	4	J48	79,7716	0,892	0,798	0,824	0,831	1,63	0,43
						RF	79,7716	0,892	0,798	0,824	0,829	46,91	2,7
						MLP	78,4233	0,89	0,784	0,813	0,861	50,22	0,22
				GainRatio	10	J48	79,7716	0,892	0,798	0,824	0,829	5,64	0,24
						RF	80,0032	0,877	0,8	0,824	0,827	75,7	2,29
						MLP	84,0601	0,9	0,841	0,858	0,891	115,27	0,34
					15	J48	79,7716	0,892	0,798	0,824	0,829	6,93	0,21
						RF	80,1224	0,841	0,801	0,817	0,801	69,5	3,68
						MLP	-	-	-	-	-	-	-
					20	J48	79,7716	0,892	0,798	0,824	0,829	5,23	0,32
						RF	-	-	-	-	-	-	-
						MLP	-	-	-	-	-	-	-
		Without	With	CfsSubsetEval	13	J48	80,4919	0,78	0,805	0,791	0,539	48,73	0,33
						RF	76,3319	0,777	0,763	0,77	0,554	228,64	4,81
						MLP	84,0387	0,763	0,84	0,791	0,584	290,49	0,32
				GainRatio	10	J48	83,0640	0,775	0,831	0,796	0,534	35,26	0,29
						RF	80,8936	0,759	0,809	0,781	0,534	203,61	7,15
						MLP	84,1124	0,771	0,841	0,794	0,526	191,88	0,26
					15	J48	75,6022	0,772	0,756	0,764	0,524	57,66	0,58
						RF	79,6096	0,777	0,796	0,786	0,527	227,01	4,47
						MLP	85,6267	0,793	0,856	0,792	0,564	324,77	0,48
					20	J48	75,9075	0,767	0,759	0,763	0,508	74,95	0,33
						RF	78,4956	0,774	0,785	0,779	0,541	268,22	3,21
						MLP	62,5343	0,752	0,625	0,675	0,511	569,87	0,57
			Without	CfsSubsetEval	7	J48	75,1162	0,745	0,751	0,748	0,485	9,02	0,18
						RF	85,4165	0,757	0,854	0,79	0,507	90,79	3,13
						MLP	56,5413	0,767	0,565	0,63	0,533	79,96	0,23
				GainRatio	10	J48	74,8805	0,747	0,749	0,748	0,502	15,13	0,53
						RF	84,2556	0,756	0,843	0,789	0,52	96,99	3,53
						MLP	85,6695	0,734	0,857	0,791	0,505	102,47	0,34
					15	J48	59,5244	0,757	0,595	0,654	0,509	17,18	0,6
						RF	59,9127	0,76	0,599	0,657	0,518	146,94	18,85
						MLP	-	-	-	-	-	-	-
					20	J48	71,5600	0,759	0,716	0,735	0,521	10,54	0,35
						RF	-	-	-	-	-	-	-
						MLP	-	-	-	-	-	-	-

ANNEX 12: TEST RESULTS TABLE OF UNDERSAMPLING TECHNIQUE

Dataset	Sampling Technique	Variable *Dep_Delay + Real_Dep_Time* Presence	Outliers Presence	Attribute Selection	Algorithm		CCI	PR	RE	F	ROC	T to build model	T to test model
FlightData	Undersampling	With	With	CfsSubsetEval	3	J48	79,7716	0,892	0,798	0,824	0,829	0,35	0,07
						RF	79,7555	0,892	0,798	0,824	0,839	7,63	1,47
						MLP	79,7622	0,892	0,798	0,824	0,888	89,88	0,2
				GainRatio	10	J48	79,7569	0,892	0,798	0,824	0,831	3,93	0,21
						RF	79,6203	0,892	0,796	0,823	0,848	39,88	2,33
						MLP	28,1736	0,86	0,282	0,283	0,892	162,72	0,35
					15	J48	79,7716	0,892	0,798	0,824	0,829	6,79	0,54
						RF	75,2045	0,885	0,752	0,787	0,835	63,68	3,2
						MLP	-	-	-	-	-	-	-
					20	J48	79,7716	0,892	0,798	0,824	0,829	751	0,36
						RF	-	-	-	-	-	-	-
						MLP	-	-	-	-	-	-	-
			Without	CfsSubsetEval	3	J48	79,7716	0,892	0,798	0,824	0,829	0,26	0,18
						RF	73,8174	0,883	0,738	0,776	0,829	5,97	2,19
						MLP	79,7716	0,892	0,798	0,824	0,839	45,32	0,3
				GainRatio	10	J48	79,7716	0,892	0,798	0,824	0,829	1,16	0,28
						RF	74,8002	0,881	0,748	0,784	0,831	27,07	2,77
						MLP	-	-	-	-	-	-	-
					15	J48	79,7716	0,892	0,789	0,824	0,829	1,46	0,32
						RF	74,8818	0,878	0,749	0,784	0,827	29,88	3,37
						MLP	-	-	-	-	-	-	-
					20	J48	79,7716	0,892	0,798	0,824	0,829	1,18	0,31
						RF	-	-	-	-	-	-	-
						MLP	-	-	-	-	-	-	-
		Without	With	CfsSubsetEval	12	J48	69,016	0,784	0,69	0,727	0,565	13,42	0,54
						RF	48,6725	0,764	0,487	0,559	0,526	103,73	3,18
						MLP	-	-	-	-	-	-	-
				GainRatio	10	J48	29,5433	0,724	0,295	0,336	0,458	3,86	0,2
						RF	20,9407	0,692	0,209	0,193	0,473	44,13	2,56
						MLP	85,6695	0,734	0,857	0,791	0,501	160,89	0,45
					15	J48	69,016	0,784	0,69	0,727	0,565	14,47	0,25
						RF	-	-	-	-	-	-	-
						MLP	-	-	-	-	-	-	-
					20	J48	47,6227	0,779	0,476	0,547	0,52	18,13	0,82
						RF	-	-	-	-	-	-	-
						MLP	-	-	-	-	-	-	-
			Without	CfsSubsetEval	6	J48	85,6695	0,734	0,857	0,791	0,5	1,52	0,17
						RF	54,5061	0,759	0,545	0,613	0,518	63,31	2,85
						MLP	-	-	-	-	-	-	-
				GainRatio	10	J48	71,7648	0,779	0,718	0,744	0,561	3	0,32
						RF	53,6974	0,76	0,537	0,606	0,519	91,07	8,63
						MLP	-	-	-	-	-	-	-
					15	J48	71,7662	0,779	0,718	0,744	0,563	4,55	0,38
						RF	-	-	-	-	-	-	-
						MLP	-	-	-	-	-	-	-
					20	J48	71,9978	0,775	0,72	0,744	0,553	2,75	0,42
						RF	-	-	-	-	-	-	-
						MLP	-	-	-	-	-	-	-